Intra-Class Random Erasing (ICRE) augmentation for audio classification[1]

Teerath Kumar *Jinbae Park **Sung-Ho Bae
Kyung Hee University
teerathkumar142@gmail.com *qkrwlsqo94@gmail.com **shbae@khu.ac.kr

Abstract

Data augmentation has been helpful in improving the performance in deep learning, when we have a limited data and random erasing is one of the augmentations that have shown impressive performance in deep learning in multiple domains. But the main issue is that sometime it loses good features when randomly selected region is erased by some random values, that does not improve performance as it should. We target that problem in way that good features should not be lost and also want random erasing at the same time. For that purpose, we introduce new augmentation technique named Intra-Class Random Erasing (ICRE) that focuses on data to learn robust features of the same class samples by randomly exchanging randomly selected region. We perform multiple experiments by using different models including resnet18, VGG16 over variety of the datasets including ESC10, UrbanSound8K. Our approach has shown effectiveness over others methods including random erasing.

1. Introduction

Deep learning has been successful in multiple domains such as computer vision, natural language processing and sound recognition [1]. In all these domain, main focus is to design the architecture of neural networks [2, 3, 4, 5]. However, these models can easily overfit the data and are data hungry. Data augmentation and regularization are the common techniques used to cope these issues. Recently, there are number of techniques used for augmentation and regularization to make model generalize such as random erasing [6], random time shift [7], frequency shift [7], Gaussian noise [7], flipping [8], dropout [9] and batch normalization [10]. In random time and frequency shift [7], there is horizontal and vertical translation to Mel Spectrogram as image, respectively. Both are considered as natural perturbation to image, in audio data case, audio data converted into Mel Spectrogram. Gaussian noise [7] is used to handle the problem, when there is possibility that noise may be recorded while transforming or recording audio. In flipping [8], Mel Spectrogram is flipped vertically and horizontally with given probability during training, that has

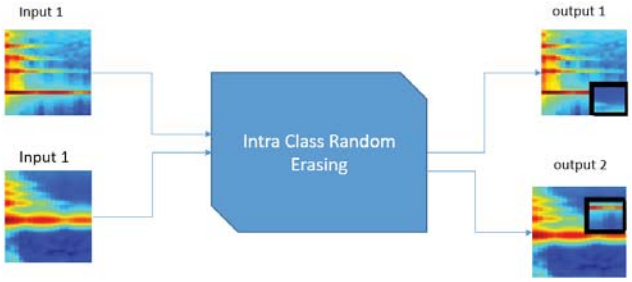

Figure 1. Architecture: Two inputs of same class are passed to ICRE augmentation, it produces two output with randomly exchange region of fix window on random position.

[1] This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2019-01-01768, Deep Neural Network based Real-Time Accurate Voice Source Localization using Drones)

shown an improvement over baseline. In dropout [9], main idea is to drop units randomly with connections during training, this is prevents neural network from overfitting.

Random erasing [6] is one of the augmentations that is widely used in deep learning for multiple domains. In random erasing [6], random region of image is selected and random value or mean value of dataset is used to replace all the values in the selected region during the training phase. But there is a problem with random erasing, sometime it loses the good features, that does not improve the performance as it should. To cope this problem, we introduce new augmentation ICRE, it enables random erasing and does not lose good feature as random erasing does.

In this paper, we deal with this problem by introducing the new ICRE augmentation that provides good features of the same class by exchanging randomly selected portion of one sample with randomly selected portion of another sample of same class. There are two benefits of doing this, firstly, good features cannot be lost and secondly, it preserves the random erasing. Exchanging randomly selected region is very important in audio domain, because Mel Spectrogram is very sensitive to noise and if we exchange randomly selected region of the same class samples, it provides same class sound but with different voice in between. From that, model is able to learn that it is different voice of same class, in this way it learns the good features and random erasing is also present at the same time.

2. Method

In this section, we define our method of ICRE. Let X be samples of class C_i, and W is window size. First to get window length and width by multiplying length and width with W. Then get random index of sample number and get random index of length and width, from which you want to crop the region. Finally exchange the region with another sample region. Repeat this process for all classes. As explained in fig 2.

2.1. Augmentations

Algorithm 1 Intra-Class-Random-Erasing-Augmentation

procedure IntraClassRandomErasing(x, $windowSize = 1/2$)
  Input : x, samples of the class C_i
  Input : WindowSize, (by default 1/2)
  Output : Return randomly exchanged samples with selected region.
  $wLength \leftarrow imageLength * WindowSize$
  $wWidth \leftarrow imageWidth * WindowSize$
  $newX \leftarrow copy(x)$
  while $i \neq len(x)/2$ **do** ▷ stop when all samples are mixed
    ▷ Random length and width index selection of window
    $randLI \leftarrow rand(0, wLength)$ ▷ Random length index
    $randWI \leftarrow rand(0, wWidth)$ ▷ Random width index
    ▷ Random index of sample other than current i sample
    $randI \leftarrow rand(i, len(x) - 1)$
    ▷ Now exchange window of two different index samples
    $temp \leftarrow copy(newX[i])$
    $IndexI = (i, randL : randL + wLength, randW : randW + wWidth)$
    $IndexR = (randI, randL : randL + wLength, randW : randW + wWidth)$
    $newX[IndexI] \leftarrow copy(x[IndexR])$
    $x[IndexR] \leftarrow copy(temp[randL : randL + wLength, randW : randW + wWidth])$
    $i = i + 1$
  return $concatenate(x_new, x), concatenate(C_i, C_i)$

Figure 2. Algorithm: IntraClass Random Erasing.

To compare with others familiar augmentations, we used below augmentations.

2.1.1. Random Erasing

In RE [6], randomly selected region is replaced with random values or mean of samples. Occlusion is generated on training samples, which prevents from over-fitting and make model more robust.

2.1.2. Random Time and Frequency Shift

Since all audio samples are converted into log Mel spectrogram. On x-axis and y-axis, there is time dimension and frequency dimension, respectively. In random time and frequency shift [7], there is random shift in time and frequency dimension. We use 0.01 percentage of original as random shift.

2.1.3. Gaussian Noise

Gaussian noise [7] is considered as perturbation to add noise and to prevent model from over-fitting. It is especially useful, where noise is recorded while recording or transforming the audio data.

3. Experimental and Results

In this section, we discuss detail of experiments.

3.1. Datasets

For checking the effectiveness of our approach we used two popular datasets i.e ESC10, UrbanSound8K.

3.1.1. ESC10

ESC-10[11] is the environmental sound classification dataset. It consists of 10 classes (dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster, fire crackling) and 400 audio files. The length of each audio is 5 seconds, and sampled at 44.1kHz. We extract the log Mel spectrogram and resize it to 128×216. We randomly split the data into 0.8: 0.2 for training and testing

3.1.2 UrbanSound8K

UrbanSound8K[3] consists of 8732 labeled sound samples less than 4 seconds each. There are total 10 classes (air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music).

We extract the log Mel spectrogram and resize it to 32×32. There are total 10 folds for better comparison. As there are 10 folds, 9 for training and 1 for testing. It means we have to do 10 different experiments.

3.2. Experimental setting

We use VGG16[12] and resnet18[13] models, with stochastic gradient descent, learning rate of 0.1 and 150 epochs. For ESC10 dataset, we train each model three times and results are reported using confidence interval. For UrbanSound8K dataset, as it is 10 fold dataset, we use 10 fold cross validation and average results are reported for each model.

3.3. Results

We compare results with other augmentations i.e horizontal and vertical shifting, Gaussian noise as noise addition, random erasing.

Resnet18 – ESC10	
Augmentation	**Accuracy**
No Augmentation	79.77 ± 3.42
Horizontal & Vertical Shifting	82.03± 3.16
Gaussian Noise	81.30± 2.14
Random Erasing	82.63±1.03
ICRE (our)	**86.49±3.29**

Table 1. Resnet18 results for ESC10 dataset

Resnet18 – UrbanSound 8K	
Augmentation	**Accuracy**
No Augmentation	74.004
Horizontal & Vertical Shifting	75.67
Gaussian Noise	74.59
Random Erasing	74.17
ICRE (our)	**76.56**

Table 2. Resnet18 results for UrbanSound 8K dataset

VGG16 – ESC10	
Augmentation	**Accuracy**
No Augmentation	72.15± 4.45
Horizontal & Vertical Shifting	76.47± 5.96
Gaussian Noise	78.33±3.58
Random Erasing	74.92±12.44
ICRE (our)	**83.37± 4.5**

Table 3. VGG16 results for ESC10 dataset

VGG16 – UrbanSound 8K	
Augmentation	**Accuracy**
No Augmentation	71.42
Horizontal & Vertical Shifting	72.93
Gaussian Noise	72.94
Random Erasing	72.23
ICRE (our)	**72.45**

Table 4. VGG16 results for UrbanSound 8K dataset

For resnet18 ESC10 (table 1) dataset and UrbanSound8K dataset (table 2), our method outperforms other related augmentation methods. Further checking the effectiveness of our method, we use another model VGG16 for both datasets (table 3 and table 4), it outperforms the other augmentation methods.

4. Conclusion

We found issue with random erasing, since good features may be erased, reason that accuracy not improved as expected. To address this issue, we introduced the new augmentation, Intra Class Random Erasing (ICRE). For checking the effectiveness of our method, we used different models over variety of datasets and compared with other augmentations, our method outperforms the other augmentations methods. Our future direction is to apply ICRE on image domain.

References

[1] Gil Keren, Jun Deng, Jouni Pohjalainen, and Bjorn W ¨ Schuller, "Convolutional neural networks with data augmentation for classifying speakers' native language.," in INTERSPEECH, 2016, pp. 2393–2397.

[2] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for lvcsr," in 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013, pp. 8614–8618.

[3] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in International conference on machine learning, 2014, pp. 1764–1772.

[4] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4960–4964.

[5] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016, pp. 4945–4949.

[6] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang, "Random erasing data augmentation," in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020.

[7] Kangkang Lu, Chuan-Sheng Foo, Kah Kuan Teh, Huy Dat Tran, and Vijay Ramaseshan Chandrasekhar, "Semi-supervised audio classification with consistencybased regularization.," in INTERSPEECH, 2019, pp. 3654–3658.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[9] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[10] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[11] [11] Karol J Piczak, "Esc: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[14] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.