

# Normalization Before Shaking Toward Learning Symmetrically Distributed Representation Without Margin in Speech Emotion Recognition

Che-Wei Huang, *Student Member, IEEE*, and Shrikanth S Narayanan, *Fellow, IEEE*

**Abstract**—Regularization is crucial to the success of many practical deep learning models, in particular in a more often than not scenario where there are only a few to a moderate number of accessible training samples. In addition to weight decay, data augmentation and dropout, regularization based on multi-branch architectures, such as Shake-Shake regularization, has been proven successful in many applications and attracted more and more attention. However, beyond model-based representation augmentation, it is unclear how Shake-Shake regularization helps to provide further improvement on classification tasks, let alone the baffling interaction between batch normalization and shaking. In this work, we present our investigation on Shake-Shake regularization, drawing connections to the vicinal risk minimization principle and discriminative feature learning in verification tasks. Furthermore, we identify a strong resemblance between batch normalized residual blocks and batch normalized recurrent neural networks, where both of them share a similar convergence behavior, which could be mitigated by a proper initialization of batch normalization. Based on the findings, our experiments on speech emotion recognition demonstrate simultaneously an improvement on the classification accuracy and a reduction on the generalization gap both with statistical significance.

**Index Terms**—Shake-Shake Regularization, Batch Normalization, Symmetrically Distributed Representation, Vicinal Risk Minimization, Prediction Uncertainty Minimization, Affective Computing, Speech Emotion Recognition

## 1 INTRODUCTION

Deep convolutional neural networks have been successfully applied to several pattern recognition tasks such as image recognition [1], machine translation [2] and speech emotion recognition [3]. Currently, to successfully train a deep neural network, one needs either a sufficient number of training samples to implicitly regularize the learning process, or employ techniques like weight decay and dropout [4] and its variants to explicitly keep the model from overfitting.

In the recent years, one of the most popular and successful architectures is the residual neural network (ResNet) [1] and its variant ResNeXt [5] with multiple residual branches. The ResNet architecture was designed based on a key assumption that it is more efficient to optimize the residual term than the original task mapping. Since then, a great deal of effort in machine learning and computer vision has been dedicated to study the multi-branch architecture.

Deep convolutional neural networks have also gained much attention in the community of affective computing mainly because of its outstanding ability to formulate discriminative features for the top-layer classifier. Usually the number of parameters in a model is far more than the number of training samples and thus it requires heavy regularization to train deep neural networks for affective computing. However, since the introduction of batch nor-

malization [6], the gains obtained by using dropout for regularization have decreased [6], [7], [8]. A recent work dedicated to study the disharmony between dropout and batch normalization [9] suggests that dropout introduces a variance shift between training and testing, which causes batch normalization to malfunction if batch normalization is placed after dropout, which severely limits the application of successful architectures such as ResNet or the application of dropout to the top-most fully connected layers. Yet, multi-branch architectures have emerged as a promising alternative for regularizing convolutional layers.

Regularization techniques based on multi-branch architectures such as Shake-Shake [10] and ShakeDrop [11] have delivered impressive performances on standard image datasets such as the CIFAR-10 [12] dataset. In a clever way, both of them utilize multiple branches to learn different aspects of the relevant information and then a summation in the end follows for information alignment among branches. Also, both of Shake-Shake and ShakeDrop regularizations emphasize on the important interaction between batch normalization and shaking. However, none of them gave an explanation for this phenomenon, other than a brief discussion on limiting the strength of shaking. Instead of using multiple branches, a recent work [13] based on a mixture of experts showed that randomly projecting samples is able to break the structure of adversarial noise that could easily confound the model and as a result mislead the learning process. Despite not being an end-to-end approach, it shares the same idea of integrating multiple streams of model-based diversity.

In addition, a recent trend of studies on data augmentation, based on the Vicinal Risk Minimization (VRM)

• The authors are with the Signal Analysis and Interpretation Laboratory (SAIL), Department of Electrical Engineering and Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089 USA

E-mail: cheweihu@usc.edu; shri@sipi.usc.edu

• This work is supported by NSF.

[14] principle, proposed to interpolate and/or extrapolate training samples in feature space, for example, [15] and [16]. Szegedy et al. [17] studied regularization of training by label smoothing. Furthermore, Mixup [18] performed convex combinations of pairs of feature and label to demonstrate that expanding the coverage of training samples in feature space, leaving little to none of margin between classes, could not only improve the performance of a model but also make it robust to adversarial samples. Effectively, Mixup reduces the uncertainty in prediction of a testing sample lying outside of the coverage of training samples in feature space by linear interpolation. Based on Mixup, Manifold Mixup [19] called for mixing intermediate representations instead of raw inputs.

Our work follows the model-based representation augmentation thread like Shake-Shake, ShakeDrop regularization and Manifold Mixup. In this work, we study the Shake-Shake regularized ResNeXt for speech emotion recognition. In addition to shaking the entire spectral-temporal feature maps with the same strength, we propose to address different spectral sub-bands independently based on our hypothesis of the non-uniform distribution of affective information over the spectral axis. Furthermore, we investigate and come up with an explanation for the ability of shaking regularization to improve classification tasks and its crucial interaction with batch normalization. In order to achieve our goal, we conduct ablation studies on MNIST [20] and CIFAR-10 datasets to highlight a subtle difference in the requirement of optimal embeddings by classification tasks based on the VRM principle and by verification tasks. In addition, we identify a strong resemblance in the mathematical formulation between batch normalized residual blocks and batch normalized recurrent neural networks, where both of them suffer from a shared issue: *faster convergence but more over-fitting* and could be fixed by the same technique.

Our contributions are multi-fold. First, our work explains with visualization the key factor to the success of shaking regularization and the crucial property that batch normalization plays in a shaking regularized architecture, drawing a connection between shaking and the VRM principle, and between batch normalization and discriminative embedding learning. Second, to the best of our knowledge, our work is the first to highlight the subtle difference in the requirement of optimal embeddings by classification tasks and by verification tasks. Third, our work identify the resemblance between batch normalized residual blocks and batch normalized recurrent neural networks, and the shared issue they have. Based on the solution to batch normalized recurrent neural networks, we demonstrate a significant reduction on the generalization gap, i.e. reduced overfitting, in a batch-normalized shaking-regularized ResNeXt for speech emotion recognition without sacrificing the validation accuracy.

The outline of this paper is as follows. We review related work in the next section, including Shake-Shake regularization and its variants, discriminative feature learning in verification tasks and the vicinal risk minimization principle. Section 3 introduces sub-band shaking. Section 4 presents batch normalized shaking, including ablation studies on MNIST and CIFAR-10 datasets, and the identification of batch normalized residual blocks with batch normalized re-

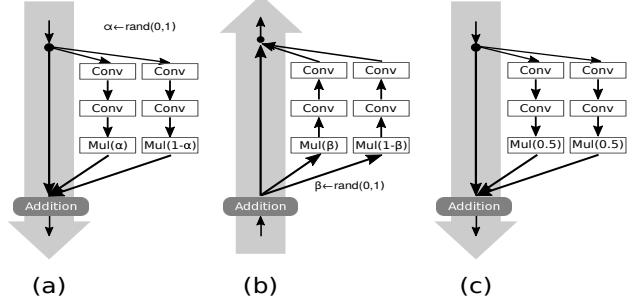


Fig. 1: An overview of a 3-branch Shake-Shake regularized residual block. (a) Forward propagation during the training phase (b) Backward propagation during the training phase (c) Testing phase. The coefficients  $\alpha$  and  $\beta$  are sampled from the uniform distribution over  $[0, 1]$  to scale down the forward and backward flows during the training phase.

current neural networks. Section 5 and 6 cover the datasets, the network architecture, the experimental setup and the results for speech emotion recognition. Section 7 concludes our work.

## 2 RELATED WORK

In this section, we review some of the recently developed techniques on which we base our work and those that are related to our findings. The first subsection covers Shake-Shake regularization [10] and its variants, including Subband Shaking [22], ShakeDrop [11] and Stochastic Shake-Shake [23] regularization. Another subsection is devoted to end-to-end discriminative feature learning algorithms in face verification, in particular the thread of work that focus on the large-margin and symmetrical representation learning in [24], [25], [26], [27], [28]. In the last subsection, we talk about a recent success in supervised learning based on the Vicinal Risk Minimization (VRM) principle [14], called Mixup [18]. Moreover, we highlight one subtle but definite difference in the desired quality of intermediate representations between classification and verification tasks.

### 2.1 Shake-Shake Regularization and Its Variants

Shake-Shake regularization is a recently proposed technique to regularize training of deep convolutional neural networks for image recognition tasks. This regularization technique based on multi-branch architectures promotes stochastic mixtures of forward and backward propagations from network branches in order to create a flow of model-based adversarial learning samples/gradients during the training phase. Owing to its excellent ability to combat overfitting even in the presence of batch normalization, the Shake-Shake regularized 3-branch residual neural network [10] has achieved one of current state-of-the-art performances on the CIFAR-10 image dataset.

An overview of a 3-branch Shake-Shake regularized ResNeXt is depicted in Fig. 1. Shake-Shake regularization adds to the aggregate of the output of each branch an addi-

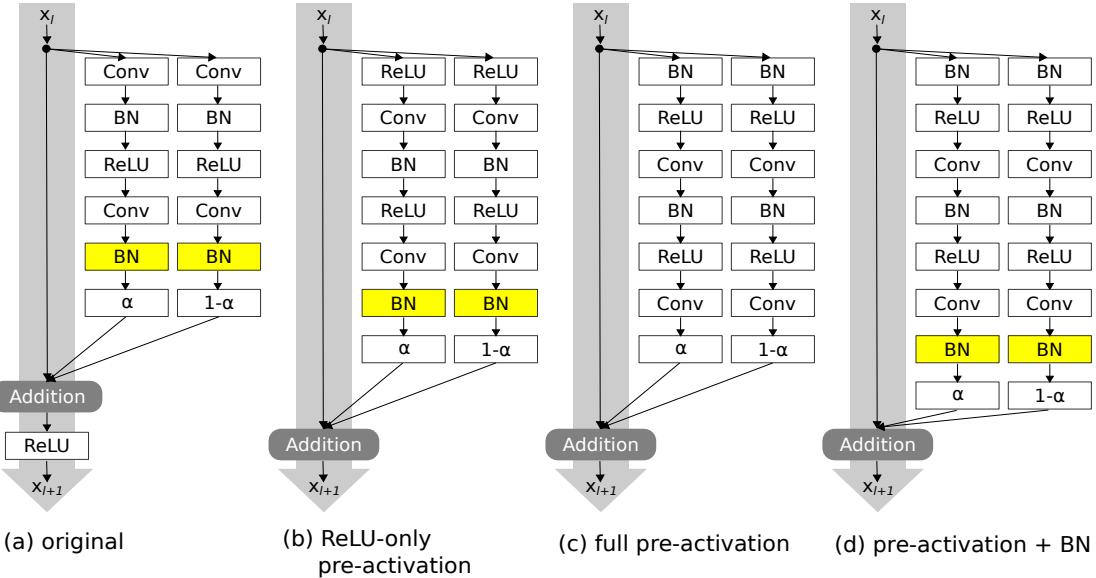


Fig. 2: Shaking regularized ResNeXt architectures with different layouts introduced in [21]

tional layer, called the shaking layer, to randomly generate adversarial flows in the following way:

$$\text{ResBlock}^N(\mathbf{X}) = \mathbf{X} + \sum_{n=1}^N \text{Shaking} \left( \{\mathbf{B}_n(\mathbf{X})\}_{n=1}^N \right)$$

where in the forward propagation for  $\mathbf{a} = [\alpha_1, \dots, \alpha_N]$  sampled from the  $(N-1)$ -simplex (Fig. 1 (a))

$$\text{ResBlock}^N(\mathbf{X}) = \mathbf{X} + \sum_{n=1}^N \alpha_n \mathbf{B}_n(\mathbf{X}),$$

while in the backward propagation for  $\mathbf{b} = [\beta_1, \dots, \beta_N]$  sampled from the  $(N-1)$ -simplex and  $\mathbf{g}$  the gradient from the top layer, the gradient entering into  $\mathbf{B}_n(x)$  is  $\beta_n \mathbf{g}$  (Fig. 1 (b)). At testing time, the expected model is then evaluated for inference by taking the expectation of the random sources in the architecture (Fig. 1 (c)). In each mini-batch, to apply scaling coefficients  $\alpha$  or  $\beta$  either on the entire mini-batch or on each individual sample independently can also make a difference.

Instead of the commonly known original or the fully pre-activation residual block, Shake-Shake regularization is proposed with the ReLU-only pre-activation residual block. Refer to Fig. 2 for more details. In addition, it has been shown that shaking with the absence of both batch normalization layers could cause the training process to diverge. One proposed remedy to this situation in [10] is to employ a shallower architecture and more importantly to reduce the range of values  $\alpha$  can take on, i.e. to reduce the strength of shaking.

ShakeDrop regularization [11], based on the same idea of model-based representation augmentation but on the deep pyramidal residual architecture, reached an improved accuracy on CIFAR-10/100 datasets. The authors further empirically observed that each residual block should end with batch normalization before the shaking layer to prevent training from diverging.

In our previous work on acoustic sub-band shaking [22] and stochastic Shake-Shake regularization [23] for affective

computing from speech, we found that in a fully pre-activation architecture without a batch normalization layer right before shaking, the shaking mechanism contributes much more to constraining the learning process than to boosting the generalization power. In addition, we showed methods to relax or control the impact of shaking, either deterministically by sub-band shaking or probabilistically by randomly turning off shaking, in order to trade-off between a higher accuracy and reduced over-fitting. Fortunately, with appropriate hyper-parameters, we could achieve both with statistical significance, compared to the baseline.

All these findings indicated there is a close interaction between shaking and batch normalization. However, these studies did not give an explanation for the crucial location of batch normalization in a shaking regularized architecture, other than a brief discussion of the range of  $\alpha$ .

## 2.2 End-to-End Discriminative Feature Learning

Recently, there has been a trend to focus on the design of loss functions so that a neural network supervised by such a loss function is able to formulate more discriminative features, usually compared to the ordinary softmax loss, for face verification. Inspired by the contrastive loss [29] and the triplet loss [30], the main design objective aims to simultaneously minimize intra-class dispersion and maximize inter-class margin. However, using the contrastive loss or the triplets loss often involves training with pairs and triplets on the order of  $O(N^2)$  and  $O(N^3)$  when  $N$  is the number of training samples, or one has to rely on a carefully selected sampling strategy.

A series of work hence reviewed the interpretation of the softmax loss as a normalized exponential of inner products between feature vector and class center vectors, and came up with various modifications for achieving the aforementioned design objective, including the large-margin softmax [24], SphereFace [25], CosFace [26], ArcFace [27] and Centralized Coordinate Learning (CCL) [28].

Different from the other four modifications, CCL distributes features dispersedly by centralizing the features to the origin of the space during the learning process so that feature vectors from different classes can be more separable in terms of a large angle between neighboring classes, and ideally symmetrically distributed in the whole feature space. The CCL loss is presented as follows:

$$\mathcal{L} = \frac{-1}{N} \sum_i^N \log \frac{e^{\Phi(\mathbf{x}_i) \cos(\theta_{y_i})}}{\sum_j^K e^{\Phi(\mathbf{x}_i) \cos(\theta_{y_j})}} \quad (1)$$

where

$$\Phi(\mathbf{x}_i)_j = \frac{\mathbf{x}_{ij} - \mathbf{o}_j}{\sigma_j}. \quad (2)$$

$\Phi(\mathbf{x}_i)_j$  and  $\mathbf{x}_{ij}$  are  $j$ -th coordinate of  $\Phi(\mathbf{x}_i)$  and  $\mathbf{x}_i$ , respectively, and  $\theta_{y_j}$  is the angle between  $\mathbf{x}_i$  and the class center vector  $\mathbf{w}_{y_j}$ . It is immediately clear that Eq. (2) resembles the famous batch normalization:

$$\Phi(\mathbf{x}_i)_j = \gamma_j \frac{\mathbf{x}_{ij} - \mathbf{o}_j}{\sigma_j} + \beta_j \quad (3)$$

except that the trainable affine transformation, defined by  $\gamma$  and  $\beta$ , after the normalization are missing in the formulation. In Eq. (2) and (3),  $\sigma$  and  $\mathbf{o}$  are running standard deviation and running mean updated per mini-batch during training. In fact, it was showed that there would be slight degradation in performance when the trainable affine transformation is employed. In the rest of this work, we use batch normalization or CCL interchangeably to refer to this discriminative feature learning.

### 2.3 Vicinal Risk Minimization and Mixup

In statistical learning theory, because the data distribution  $\mathcal{P}(x, y)$  is unknown in most practical applications, one may approximate it by the empirical distribution

$$\mathcal{P}_\delta(x, y) = \frac{1}{N} \sum_n \delta(x = x_n, y = y_n), \quad (4)$$

where  $\delta(x = x_n, y = y_n)$  is a Dirac centered at  $(x_n, y_n)$ , and  $\mathcal{D} = \{(x_n, y_n) : 1 \leq n \leq N\}$  is a given training set sampled from  $\mathcal{P}(x, y)$ . This paradigm of learning via minimizing a risk based on the approximation in Eq. (4) is referred to as the Empirical Risk Minimization (ERM) [31].

Instead of a Dirac function, one may employ a vicinity function to take into consideration the vicinity of a true training pair  $(x_n, y_n)$ , and this alternative is therefore called the Vicinal Risk Minimization (VRM) principle [14]. The Gaussian vicinity [14]  $\nu(x, y|x_n, y_n) = \mathcal{N}(x - x_n, \sigma^2)\delta(y - y_n)$  is one of the well-known examples, which is equivalent to data augmentation with additive Gaussian noise.

Recently, a novel vicinity function, called Mixup [18], is proposed to cover the entire convex hull of training pairs:

$$\nu(x, y|x_n, y_n) = \frac{1}{N} \sum_m \mathbb{E}_\lambda [\delta(x = u_m, y = v_m)], \quad (5)$$

$$u_m = \lambda \cdot x_n + (1 - \lambda) \cdot x_m, \quad (6)$$

$$v_m = \lambda \cdot y_n + (1 - \lambda) \cdot y_m, \quad (7)$$

where  $\lambda \in [0, 1]$ . This data augmentation scheme by Mixup is believed to reduce the amount of undesirable oscillations when predicting outside the original training examples as

it fills in the feature space between classes by convex combinations of true training pairs, and hence contribute both to improvement on classification accuracy as well as robustness to adversarial examples.

**Classification versus Verification:** One could easily observe that the directions for improving classification tasks based on VRM and for refining embeddings in verification tasks based on the aforementioned design objective are drastically different, where the former strives to minimize or eliminate the margin between classes while the latter aims to ensure a definite minimum margin. We provide more details in Section 4.

## 3 SUB-BAND SHAKING

Shake-Shake regularization delivers different results depending on the strength of the shaking, in terms of the range of values  $\alpha$  (or  $\beta$ ) takes on as well as whether the same  $\alpha$  (or  $\beta$ ) is shared within a mini-batch. In addition to batch- or sample-wise shaking, when it comes to the area of acoustic processing, there is another orthogonal dimension to consider: the spectral domain. Leveraging domain knowledge, our first proposed models are based on a simple but plausible hypothesis that affective information is distributed non-uniformly over the spectral axis [32]. Therefore, there is no reason to enforce the entire spectral axis to be shaken with the same strength concurrently. Furthermore, adversarial noise may exist and extend over the spectral axis. By deliberately shaking spectral sub-bands independently, the structure of adversarial noise may be broken and become less confounding to the model.

There has been work on multi-stream framework in speech processing. For example, Mallidi et al. [33] designed a robust speech recognition system using multiple streams, each of them attending to a different part of the feature space, to fight against noise. However, lacking both multiple branches and the final information alignment, the design philosophy is fundamentally different from that of multi-branch architectures. In fact, sub-band shaking could be viewed as a bridge between the multi-stream framework and the multi-branch architecture.

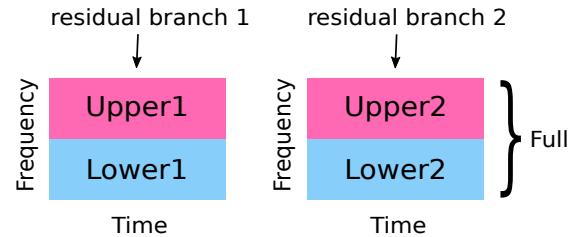


Fig. 3: An illustration for the sub-band definitions

Before we formally define the proposed models, we introduce the definition of sub-bands first. Fig. 3 depicts the definition for sub-bands in a 3-branch residual block. Here we slightly abuse the notations of frequency and time because after two convolutional layers these axes are not exactly the same as those of input to the branches; however, since convolution is a local operation they still hold the corresponding spectral and temporal nature. At the output of each branch, we define the high-frequency half to be

the upper sub-band while the low-frequency half to be the lower sub-band. We take the middle point on the spectral axis to be the border line for simplicity. The entire output is called the full band.

Having defined these concepts, we denote  $\mathbf{X}$  the input to a residual block,  $\mathbf{X}^i$  the full band from the  $i$ -th branch,  $\mathbf{X}_u^i$  the upper sub-band from the  $i$ -th branch and  $\mathbf{X}_l^i$  the lower sub-band from the  $i$ -th branch. Naturally, the relationship between them is given by  $\mathbf{X}^i = [\mathbf{X}_u^i | \mathbf{X}_l^i]$ . We also denote  $\mathbf{Y}$  the output of a Shake-Shake regularized residual block.

To study the effectiveness of sub-band shaking for speech emotion recognition, we propose the following models for benchmarking:

- 1) Shake the full band (**Full**)

$$\mathbf{Y} = \mathbf{X} + \sum_{n=1}^N \text{ShakeShake}(\{\mathbf{X}^n\}_{n=1}^N). \quad (8)$$

- 2) Shake both sub-bands but independently (**Both**)

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} + [\mathbf{Y}_u | \mathbf{Y}_l], \\ \mathbf{Y}_u &= \sum_{n=1}^N \text{ShakeShake}(\{\mathbf{X}_u^n\}_{n=1}^N), \\ \mathbf{Y}_l &= \sum_{n=1}^N \text{ShakeShake}(\{\mathbf{X}_l^n\}_{n=1}^N). \end{aligned} \quad (9)$$

## 4 BATCH NORMALIZED SHAKING

It has been shown that without batch normalization in a residual block, the shaking operation could easily cause training to diverge [10]. Even with various methods to reduce the strength of shaking such as to limit the range of values  $\alpha$  can take on or to decrease the number of residual blocks and hence the number of shaking layers, applying shaking regularization without batch normalization leads to a much inferior model, compared to a unregularized ResNeXt model with batch normalization.

**Batch Normalization for Symmetrically Distributed Representations:** From the perspective of discriminative feature learning, one may assume that without batch normalization, feature vectors from different classes may lie close to each other in an uneven fashion. In this situation, the appropriate strength of shaking may be limited by the margin of two closest classes as in the reported experiments in [10]. In fact, a recent study [34] refuted the commonly believed view on the role of batch normalization in reducing the so-called internal covariate shift. Instead, they pointed out batch normalization actually helps to significantly smooth the optimization landscape during training. Intuitively, it is expectedly easier to explore the neighborhood of well-separated class center vectors, compared to a set of unevenly distributed class center vectors, since it is less likely to have overlapping embeddings from different classes in the former case. Without batch normalization, any inappropriately large strength of shaking may lead to *overshot* embeddings, cause two or more classes to overlap and most importantly result in a slower and inferior convergence. Empirically, we find it is the key factor to keep a distribution of dispersed representations that spans over the entire feature space in a symmetric fashion and extends

away from the origin, even when shaking is applied. Therefore, we hypothesize that the batch normalization (or CCL) layer right before shaking serves to disperse embeddings and therefore prevent them from overlapping each other under the influence of perturbation that is encouraged by shaking around class center vectors.

**Model-Based Representation Augmentation for Coverage of Between-Class Feature Space:** From the review in Section 2, we have noticed that classification tasks based on the VRM principle and verification tasks ask for subtly different distributions of representations. The success of Mixup in classification tasks advocates for minimizing or eliminating the gap between classes in feature space by augmented training samples. By doing so, the chance of predicting samples lying outside of the coverage of the original or augmented training samples, which leaves room for attacks by adversarial samples [35] or results in uncertain predictions, is reduced. On the other hand, CCL, and the other discriminative feature learning algorithms for verification, could leave a larger margin between classes (for example, c.f. Fig. 4 (a) and (c)). We hypothesize that shaking following a batch normalization layer expands the coverage of training samples and effectively it eliminates the margin between classes in feature space to achieve a similar distribution as in Mixup. One caveat is that shaking does not smooth labels during exploration, and therefore is more similar to a prior work of Mixup, where interpolation and extrapolation of the nearest neighbors of the same class in feature space is proposed to enhance generalization [15]. However, shaking operates solely on a single sample at a time, unlike [18] and [15].

### 4.1 Embedding Learning on MNIST and CIFAR-10

To demonstrate the close interaction between shaking and batch normalization in representation learning, we present two ablation studies, one on the MNIST dataset and the other on the CIFAR-10 dataset. For both sets of experiments, we employ the ordinary softmax loss to examine the effectiveness of batch normalization in representation learning when it is not coupled with the softmax function.

The first set aims to visualize the embeddings of handwritten digit images learned under the influence of batch normalization and shaking. We employ a ResNeXt (20,  $2 \times 4d$ ) architecture, where the last residual block reduces the feature dimension to 2 for the purpose of visualization. Fig. 4 depicts embeddings by four layouts of residual blocks, where the top and bottom rows correspond to embeddings of the training and testing samples, respectively. From left to right, the columns represent embeddings learned by models of fully pre-activation (Fig. 2(c) **PreAct**) without shaking, **PreAct** with shaking, fully pre-activation + BN (Fig. 2(d) **PreActBN**) without shaking and **PreActBN** with shaking, respectively.

The first column serves as the baseline in this set of experiments. Immediately, we can observe a severe degradation in separability when applying shaking to **PreAct**, comparing Fig. 4(a,e) with Fig. 4(b,f). Also notice that shaking without a directly preceding batch normalization could perturb or destroy the symmetric distribution, which is obvious when there is no shaking (the symmetry in Fig.

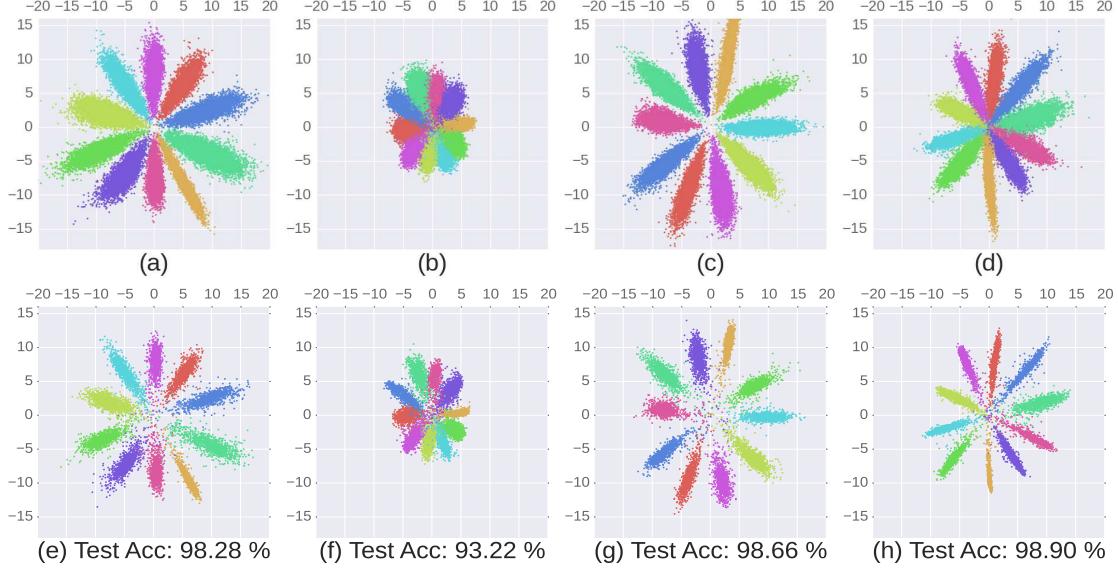


Fig. 4: MNIST embeddings based on different layouts of residual blocks. We set the feature dimension entering into the output layer to be two and train them in an end-to-end fashion. The top and bottom rows depict embeddings of the training samples extracted in the train mode (i.e.  $\alpha \in [0, 1]$ ) without updating parameters, and testing samples extracted in the eval mode ( $\alpha = 0.5$ ), respectively. (a,e) fully pre-activation (Fig. 2(c)) without shaking (b,f) fully pre-activation (Fig. 2(c)) with shaking (c,g) fully pre-activation + BN (Fig. 2(d)) without shaking (d,h) fully pre-activation + BN (Fig. 2(d)) with shaking

4(a,e)). This is rather interesting as batch normalization still exists in **PreAct** residual block, only not directly connected to the shaking layer. It seems the exploration encouraged by shaking around each class center has expanded its coverage but without a directly preceding batch normalization to maintain a good dispersion between classes, each class only expands to overlap with neighboring classes, and the resulting distribution is heavily tiled. Consequently, **PreAct** with shaking delivers a much inferior performance compared to **PreAct**.

The comparison between **PreAct** (Fig. 4(a,e)) and **PreActBN** (Fig. 4(c,g)), both without shaking, demonstrates the effectiveness of CCL in discriminative feature learning although it is not coupled with the loss function. In Fig. 4(c), not only does it maintain a symmetric distribution of classes, but also it encourages each class to expand outward and to leave more margin between neighboring classes. As a result, **PreActBN** without shaking is able to reach a better performance compared to the baseline.

Finally, **PreActBN** with shaking (Fig. 4(d,h)), on the contrary, does not lead to a larger margin between classes as **PreActBN** without shaking does. On the other hand, it seems that the shaking has expanded the coverage of each class so that all of them are directly adjacent to each other with a minimal or zero margin. We could also observe that, although batch normalization tries to maintain a symmetric distribution of feature vectors, some of the classes in Fig. 4(d,h) are slightly tilted around the outer most region. However, the most salient difference is the distribution of testing samples, where each class becomes more compact. The performance of **PreActBN** is therefore the highest among all of the models.

Based on the VRM principle, the region close to boundaries between classes in feature space is covered by aug-

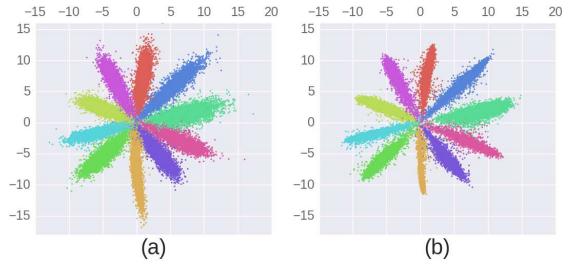


Fig. 5: Embeddings of training samples extracted in the (a) train (b) eval mode from **PreActBN** with shaking

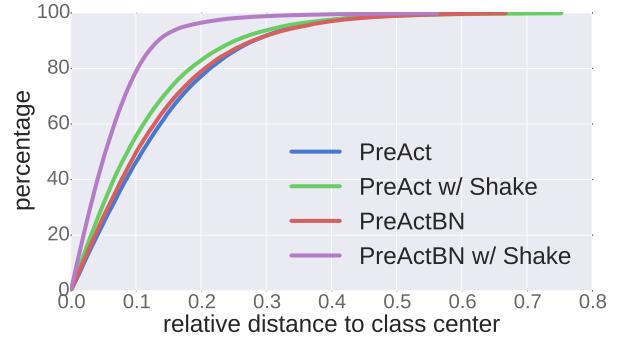


Fig. 6: Percentage of class samples within in a relative distance to class center

mented training samples. Since in supervised learning we assume testing samples are drawn from a similar or the same distribution as training samples, the majority of testing samples are mapped to embeddings close to the center of class, where most original training samples are mapped to,

and thus the distribution seems more compact. In Fig. 5, from the comparison of training embeddings in the train and eval modes, it is visually clear that shaking is expanding the coverage of training embeddings. To quantitatively show that this is the case, for each layout we calculate distances of the original training embeddings to their respective class center vectors in the eval mode and plot the percentage of class samples within distances relative to the largest distance in the class that is calculated in the train mode, in Fig. 6. It is clear that with shaking most of the original training embeddings are concentrated around the class center. For example, almost 100% of original training embeddings lies within  $0.3 \times$  largest distance for **PreActBN** with shaking and just a small number of original training samples are lying close to the boundaries.

Model	Depth	Params	Error (%)
ResNeXt (29, $16 \times 64d$ ) [5]	29	68.1M	3.58
Wide ResNet [7]	28	36.5M	3.80
Shake-Shake (26, $2 \times 96d$ ) [10]	26	26.2M	2.86
Shake-Shake (26, $2 \times 64d$ ) [10]	26	11.7M	2.98
ResNeXt (26, $2 \times 64d$ ) [10]	26	11.7M	3.76
PreActBN (26, $2 \times 64d$ ) <sup>†</sup>	26	11.7M	*2.95
BN-Shake (26, $2 \times 64d$ )	26	11.7M	*3.65
PreAct (26, $2 \times 64d$ ) <sup>†</sup>	26	11.7M	*6.92

\* average over three runs

<sup>†</sup> with shaking

TABLE 1: Test error (%) and model size on CIFAR-10

The second set of experiments on CIFAR-10 is designed to measure the contribution of the last batch normalization in **PreActBN** with shaking. In order to do so, we remove the first two batch normalization from the **PreActBN** residual block and rename the new one, the **BN-Shake** residual block (ReLU-Conv-ReLu-Conv-BN-Mul), assuming shaking is applied. Along with **PreActBN** and **PreAct**, by presenting **BN-Shake**, all of them with shaking, we are able to quantitatively demonstrate the crucial location of batch normalization in a shaking regularized architecture.

We modify the open-sourced Torch-based Shake-Shake implementation<sup>1</sup> that was released with [10] to build these three architectures. All of the rest of parameters such as the cosine learning rate scheduling and the number of epochs remain unchanged. Only the part that involves residual block definition is modified to serve our need. We run each experiment for three times to obtain a robust estimate of the performance using different random seeds. Table 1 presents the results of our experiments as well as the quoted performances on CIFAR-10 from [10].

Note that Shake-Shake ResNeXt in [10] is based on the ReLu-only pre-activation residual block (Fig. 2(b) **RPreAct**) and is thus different from **PreAct** we have here. Although the ResNeXt-26  $2 \times 64d$  is based on the **RPreAct** structure, with a shallow depth of 26, it should be comparable to one that is based on the **PreAct** when no shaking is applied [21]. Therefore, we also take it as the baseline for the pre-activation layout.

The performance of **PreActBN** with shaking (2.95%, mean of 2.89%, 3.00% and 2.95%) is comparable to the reported performance of **RPreAct** with shaking (2.98%),

1. <https://github.com/xgastaldi/shake-shake>.

where both of them have a directly preceding batch normalization layer before shaking. As expected, the performance of **PreAct** with shaking (6.92%, mean of 6.76%, 6.82% and 7.19%) is much worse than every model in Table 1, including the baseline ResNeXt-26  $2 \times 64d$ .

On the other hand, the result of **BN-Shake** (3.65%, mean of 3.56%, 3.76% and 3.62%) is rather positive. With only one batch normalization layer, it outperforms not only **PreAct** with shaking but also the baseline ResNeXt-26  $2 \times 64d$ . This finding highlights the fact that the directly preceding batch normalization plays a crucial role in keeping a good dispersion of intermediate representations when shaking is applied to explore unseen feature space, while the dispersing effect of any other batch normalization that is separated by convolutional layers from the shaking layer is reduced or only auxiliary.

Based on these two sets of experiments, it is safe to state that in the close interaction between batch normalization and shaking, batch normalization is mainly responsible for keeping a dispersed symmetric distribution of intermediate representations from perturbation by shaking, while the shaking mechanism expands the coverage of training samples to eliminate the margin between classes by promoting stochastic convex combinations of model branches, similar to [18] and [15].

For speech emotion recognition, we conduct experiments with 3-branch ResNeXt based on the original (Fig. 2(a) **PostAct**, **RPreAct**, **PreActBN** and **PreAct** layouts with and without shaking to benchmark these architectures.

## 4.2 Relation with Batch Normalized Recurrent Neural Networks

So far, we have presented arguments and analyses of experiments to clarify the interplay of batch normalization and shaking, and drawn a connection between each of them to discriminative feature learning and VRM, respectively.

In our previous work on sub-band shaking [22] and stochastic Shake-Shake regularization [23] in affective computing, we found that shaking regularization contributes more to constraining the training process, measured in terms of the generalization gap between the accuracies of training and validation, than improving generalization when using **PreAct** with shaking, which is understandable based on previous analyses. However, when employing **PreActBN** for speech emotion recognition, we further observe another interesting behavior of the resulting models. Compared to **PreAct**, although the validation performance is improved thanks to the addition of batch normalization, the gap between training and validation accuracies also drastically increased, which suggests a significantly amount of increased over-fitting.

Architecture	Shake	Valid UA (%)	Gap (%)
PreAct	✗	61.342	7.485
	✓	62.989	-1.128
PreActBN	✗	63.407	7.410
	✓	66.194	8.348

\* average over three runs

TABLE 2: Performances of **PreAct** and **PreActBN** with and without shaking for speech emotion recognition

For example, Table 2 gives a comparison between **PreAct** and **PreActBN** with and without shaking. With shaking, **PreActBN** achieves an improvement of 3.205% on the unweighted accuracy (UA) at the cost of an increment of 9.476% on the generalization gap (Gap) between training and validation UAs from **PreAct**, simply because of the addition of batch normalization. On the other hand, it is 2.787% increment on the UA and 0.938% increment on the Gap, respectively, for **PreActBN** due to the addition of shaking. Apparently, the addition of batch normalization triggers something that causes this significant change in the generalization gap when shaking is applied.

We suspect that this behavior partially resembles the reported situation in application of batch normalization to recurrent neural networks [36], [37], where Laurent et al. [36] attempted to apply batch normalization to the recurrent formulation of temporal modeling and concluded that it leads to faster convergence but more over-fitting. To address this issue, Cooijmans et al. [37] attributed the difficulties with recurrent batch normalization to gradient vanishing due to improper initialization of the batch normalization parameters, and proposed to initialize the standard deviation parameter  $\gamma$  in Eq. (3) to a small value such as 0.1, contrary to the common practice of unit initialization. With a proper initialization, they demonstrated successful applications of batch normalized Long Short Term Memory (LSTM) networks that converge faster and generalize better.

In fact, the formulations are awfully similar in batch normalized residual blocks and in batch normalized recurrent neural networks, where both of them consist of *a batch normalized sum of multiple batch normalized streams and so on so forth*. For example, in **PreActBN** with shaking the output of the  $l$ -th block is as follows:

$$\mathbf{X}_{l+1} = \alpha \text{BN}(f_l^1(\mathbf{X}_l)) + (1 - \alpha) \text{BN}(f_l^2(\mathbf{X}_l)) + \mathbf{X}_l, \quad (10)$$

and in the  $(l + 1)$ -th block,

$$f_{l+1}^i(\mathbf{X}_{l+1}) = \text{Conv}(\text{ReLU}(\text{BN}(\text{Conv}(\mathbf{Y}^i)))), \quad (11)$$

$$\mathbf{Y}^i = \text{ReLU}(\text{BN}(\mathbf{X}_{l+1})). \quad (12)$$

On the other hand, the batch normalized LSTM is given by the following equations [37]:

$$\begin{pmatrix} \mathbf{f}_t \\ \mathbf{i}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \text{BN}(\mathbf{W}_h \mathbf{h}_{t-1}) + \text{BN}(\mathbf{W}_x \mathbf{x}_t) + \mathbf{b}, \quad (13)$$

$$\mathbf{c}_t = \sigma(\mathbf{f}_t) \odot \mathbf{c}_{t-1} + \sigma(\mathbf{i}_t) \odot \tanh(\mathbf{g}_t), \quad (14)$$

$$\mathbf{h}_t = \sigma(\mathbf{o}_t) \odot \tanh(\text{BN}(\mathbf{c}_t)), \quad (15)$$

where  $\mathbf{x}_t$  is the input,  $\mathbf{W}_x$  and  $\mathbf{W}_h$  are the input and recurrent weights,  $\mathbf{f}_t$ ,  $\mathbf{i}_t$ ,  $\mathbf{o}_t$  and  $\mathbf{g}_t$  are forget, input, output and gates and the cell candidate,  $\mathbf{c}_t$  and  $\mathbf{h}_t$  are the cell and output vectors,  $\odot$  and  $\sigma$  are the Hadamard product and the logistic sigmoid function, respectively.

In these two batch normalized formulations, we find  $\mathbf{Y}^i$  and  $\mathbf{h}_t$  rather similar to each other, up to only some component-wise scaling and clipping by activations. In addition to benchmarking different layouts of residual blocks, we also present speech emotion recognition experiments to investigate the convergence behavior of shaking regularized

ResNeXt networks with the batch normalization parameter  $\gamma$  initialized to different values.

## 5 DATABASE AND EXPERIMENTS FOR SPEECH EMOTION RECOGNITION

We use six publicly available emotion corpora to demonstrate the effectiveness of the proposed models, including the eINTERFACE'05 Audio-Visual Emotion Database [38], the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [39], the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [40], the Berlin Database of Emotional Speech (Emo-DB) [41], the EMOVO Corpus [42] and the Surrey Audio-Visual Expressed Emotion (SAVEE) [43]. Some of these corpora are multi-modal in which speech, facial expression and text all convey a certain degree of affective information. However, in this paper we solely focus on the acoustic modality for experiments.

We formulate the experimental task into a sequence classification of 4 classes, including joy, anger, sadness and fear. We perform sub-utterance sampling [44] by dividing long utterances into several short segments of 6.4-second long with the same label to limit the length of the longest utterance; for example, a 10-second long angry utterance is replaced by two angry segments corresponding to the first 6.4 and the last 6.4 seconds of the original utterance with some overlapping part. In this way, we also slightly benefit from data augmentation. As a result, we obtain 6803 emotional utterances from the aggregated corpora. Table 3 summarizes the information about these six corpora.

Corpus	No. Actors	No. Utterances			
		Joy	Anger	Sadness	Fear
eINTERFACE	42	207	210	210	210
RAVDESS	24	376	376	376	376
IEMOCAP	10	720	1355	1478	0
Emo-DB	10	71	127	66	69
EMOVO	6	84	84	84	84
SAVEE	4	60	60	60	60
Total	96	1518	2212	2274	799

TABLE 3: An overview of these selected corpora, including the number of actors and the distribution of utterances in the emotional classes

Corpus	Actor Set Partition			
	1	2	3	4
eINTERFACE	2F, 8M	2F, 9M	3F, 8M	2F, 8M
RAVDESS	3F, 3M	3F, 3M	3F, 3M	3F, 3M
IEMOCAP	1F, 2M	1F, 1M	1F, 1M	2F, 1M
Emo-DB	1F, 2M	1F, 1M	1F, 1M	2F, 1M
EMOVO	1F, 0M	1F, 1M	1F, 1M	0F, 1M
SAVEE	0F, 1M	0F, 1M	0F, 1M	0F, 1M
Total	8F, 16M	8F, 16M	9F, 15M	9F, 15M

TABLE 4: F: female, M: male. The gender and corpus distributions in each actor set partition of the cross validation

For the evaluation, we adopt a 4-fold cross validation strategy. To begin with, we split the actor set into 4 partitions. In addition, we impose extra constraints to make sure that each partition is as gender and corpus uniform as possible. For example, each actor set partition is randomly distributed with 2-3 female actors and 8-9 male actors from

Layer Name	Structure	Stride	No. Params
prelim-conv	$2 \times 16, 4$	$\times 1$	[1, 1] 132
res-8	$2 \times 16, 8$ $2 \times 16, 8$	$\times 3$	[1, 1] 22.84K
res-16	$2 \times 16, 16$ $2 \times 16, 16$	$\times 1$	[1, 1] 24.88K
res-32	$2 \times 16, 32$ $2 \times 16, 32$	$\times 1$	[1, 1] 99.17K
average	-	-	-
affine	$256 \times 4$	-	1024
total	-	-	148K

TABLE 5: ResNeXt ( $12, 2 \times 8d$ ) network architecture

the eINTERFACE’05 corpus. More details are provided in Table 4. By partitioning the actor set, it becomes easier to maintain speaker independence between training and validation throughout all of the experiments.

We extract the spectrograms of each utterance with a 25ms window for every 10ms using the Kaldi [45] library. Cepstral mean and variance normalization is then applied on the spectrogram frames per utterance. To equip each frame with a certain context, we splice it with 10 frames on the left and 5 frames on the right. Therefore, a resulting spliced frame has a resolution of  $16 \times 257$ . Since emotion involves a longer-term mental state transition, we further down-sample the frame rate by a factor of 8 to simplify and expedite the training process.

We build an architecture of ResNeXt ( $12, 2 \times 8d$ ) that consists of only 3 residual stages, res-8, res-16 and res-32, in which the filter size is fixed to  $2 \times 16$ . There are 3 residual blocks in res-8, and one residual block in res-16 and res-32. Before the affine transformation, we add a dropout layer with probability of 0.5. Table 5 contains details of the architecture. For each utterance, a simple mean pooling is taken at the output of the final residual block to form an utterance representation before it is mapped by the affine layer. We avoid explicit temporal modeling layers such as a long short-term memory recurrent network because our focus is to investigate the effectiveness by shaking the ResNeXt. Note that a shaking layer has no parameter to learn and hence the model size in this work stays almost constant and only changes slightly in the situation when the number of batch normalization layers differs, since batch normalization has a set of trainable scaling and shift parameters.

We implement the shaking layer as well as the entire network architecture using the PyTorch [46] library. Only the Shake-Shake combination [10] is used and shaking is applied independently per frame. We may also for simplicity refer to the Shake-Shake regularization as shaking regularization. Due to class imbalance in the aggregated corpora, the objective function for training is the weighted cross-entropy, where the class weight is inversely proportional to the class size. The models are learned using the Adam optimizer [47] with an initial learning rate of 0.001 and the training is carried out on an NVIDIA Tesla P100 GPU. We use a mini-batch of 8 utterances across all model training and let each experiment run for 1000 epochs, and for three runs with different random seeds to obtain a robust estimate of the performance.

For sub-band shaking experiments, we initially employ the fully pre-activation layout (Fig. 2(c)). **PreAct** denotes

the model without shaking, while **PreAct-Full** and **PreAct-Both** correspond to the models with the **Full** and **Both** shaking (Eq. 8, 9), respectively. We only consider two sub-bands, defined by the middle point on the spectral axis, to present our point. However, it does not rule out the feasibility of experiments with more sub-bands. In fact, in the multi-stream framework for speech recognition, more sub-bands were considered to generate more combinations and diversity.

We conduct experiments to benchmark different layouts, including **PostAct** (Fig. 2(a)), **RPreAct** (Fig. 2(b)) and **PreActBN** (Fig. 2(d)) with and without (**Full**) shaking. Furthermore, we choose three different initialization values of  $\gamma$  in batch normalization and apply them to each of **PostAct**, **RPreAct** and **PreActBN** layouts to investigate if batch normalized ResNeXt networks also behave as batch normalized recurrent neural networks in terms of faster convergence and better generalization, in addition to the resemblance in mathematical formulations. In this set of experiments, we use the affine transformation in batch normalization despite CCL recommends not to use it, except in the last batch normalization before the output layer, where we only standardize the input and do not rescale it.

Finally, we revisit sub-band shaking with the **PreActBN** layout and with different initialization values of  $\gamma$  to present a complete comparison between shaking on the full band and on the sub-bands independently. With this set of experiments, we may also examine the effectiveness of trade-off between the un-weighted accuracy (UA) and the generalization gap (Gap) between training and validation UAs when batch normalization is applied right before shaking.

## 6 EXPERIMENTAL RESULTS

### 6.1 Sub-band Shaking

The result of sub-band shaking is presented in Table 6. It is clear that both **PreAct-Full** and **PreAct-Both** are able to improve from **PreAct**, unlike the experiments on MNIST and CIFAR-10 datasets. For both shaking regularized models, however, their improvements on the UA is comparably smaller than their reduction on the generalization gap, which suggests the models are learning a harder problem in the presence of shaking. It could also be the difference in the nature of tasks or in the number of classes that **PreAct** with shaking is improving instead of degrading.

Model	Valid UA (%)	Gap (%)
PreAct	61.342	7.485
PreAct-Full	<sup>*</sup> 62.989	<sup>†</sup> – 1.128
PreAct-Both	<sup>*</sup> 64.973	<sup>†</sup> 1.791

<sup>\*</sup> average over three runs

<sup>†</sup> significantly ( $p < 0.05$ ) outperforms **PreAct**

TABLE 6: Performances of **PreAct**, **PreAct-Full** and **PreAct-Both**

Furthermore, we conduct statistical hypothesis testing to examine the significance of the observed improvements. The resulting p-values from one-sided paired t-test indicates that the improvements of **PreAct-Full** and **PreAct-Both** from **PreAct** are statistical significant, in terms of the UA and

the Gap. In addition, compared to the improvements of **PreAct-Full**, **PreAct-Both** has reached a higher UA but also a larger gap. In fact, the improvement of **PreAct-Both** from **PreAct-Full** is also significant in terms of UA. On the other hand, **PreAct-Full** has a significant reduction on the Gap, compared to the reduction of **PreAct-Both** on the Gap. More details are presented in Table 7.

Model Pair	Valid UA	Gap
PreAct vs PreAct-Full	$1.44 \times 10^{-2}$	$2.7 \times 10^{-4}$
PreAct vs PreAct-Both	$1.22 \times 10^{-5}$	$3.0 \times 10^{-3}$
PreAct-Full vs PreAct-Both	$2.60 \times 10^{-3}$	$9.9 \times 10^{-1}$

TABLE 7: P-values resulted from one-sided paired t-test with  $df=4 \times 3 - 1 = 11$

Based on these observations, we can think of sub-band shaking as a method to relax the strength of shaking. This way, one may trade off between the amount of improvement on the UA and the amount of reduction on the Gap. Our previous work on stochastic Shake-Shake regularization [23] also investigated another method to trade off between these two performance metrics by randomly turning off shaking. Since sub-band shaking and stochastic Shake-Shake behave similarly and the latter do not contribute directly into the focus of this work, we choose to present only sub-band shaking here.

## 6.2 Experiments with Different Layouts

The results of benchmarking the layouts are presented in Table 8. First of all, we notice that even without shaking, all of **PostAct**, **RPreAct** and **PreActBN** have achieved higher UAs with significance (denoted by \*), compared to that of **PreAct**, while the UA of the first three models are comparable, i.e. no statistical significance in the differences. These improvements on the UA could be attributed to discriminative feature learning, i.e. CCL, since we additionally apply batch normalization without rescaling before the final affine layer in the first three models. We deliberately employ CCL in these three models to set up a set of competitive baselines.

Next, we look at the layouts with shaking at  $\gamma_0 = 1.0$ . We immediately observe that even with the common practice of unit initialization for  $\gamma$ , all of these three layouts give further improvements on the UA. However, none of them is able to reduce the generalization gap with statistical significance. The closest one is **RPreAct** with shaking when  $\gamma_0 = 1.0$  and the p-value for the testing is 0.088 due to a high variation of the generalization gap from each fold. With the common practice of unit initialization, **PreActBN** converges to the most accurate model with the UA as high as 66.194%, but it also results in the largest generalization gap.

In fact, we could see that for all values of  $\gamma_0$  we choose, these three layouts all outperform their respective baseline models, the same layout without shaking, in terms of UA with significance (denoted by †), and all of them show a trend of generalization gap reduction only that the majority of them are unable to reduce the generalization gap with significance, except two of them. Both of **RPreAct** and **PreActBN** outperform their baseline models simultaneously on the UA and on the Gap, achieving a higher accuracy as well as reduced over-fitting at the same time, when  $\gamma$  is

Architecture	Shake	$\gamma_0$	Valid UA (%)	Gap (%)
PreAct	✗	—	61.342	7.485
	✓	—	†62.989	† – 1.128
PostAct	✗	—	*62.782	7.822
	✓	1.00	†65.536	7.657
	✓	0.20	†65.490	5.431
	✓	0.10	†65.483	7.255
	✓	0.05	†65.512	7.205
RPreAct	✗	—	*63.939	8.517
	✓	1.00	†65.859	7.063
	✓	0.20	†65.821	8.305
	✓	0.10	†64.899	7.958
	✓	0.05	†65.052	†5.821
PreActBN	✗	—	*63.407	7.410
	✓	1.00	†66.194	8.348
	✓	0.20	†65.789	†5.817
	✓	0.10	†66.097	†6.040
	✓	0.05	†66.418	† † 3.416

\* average over three runs

\* significantly ( $p < 0.05$ ) outperforms **PreAct** without shaking

† significantly ( $p < 0.05$ ) outperforms the same layout without shaking

‡ significantly ( $p < 0.05$ ) outperforms the same layout at  $\gamma_0 = 1.0$

TABLE 8: Performances of **PreAct**, **PostAct**, **RPreAct** and **PreActBN** with and without shaking for speech emotion recognition, where  $\gamma_0$  is the initialization value of the standard deviation parameter  $\gamma$  in batch normalization

initialized as 0.05. **PreActBN** with shaking at  $\gamma_0 = 0.05$  is also the best model, both highest on the UA and lowest on the Gap (except for **PreAct** with shaking, which is known to be difficult to learn). In other words, the result of this model validates our hypothesis with statistical significance that batch normalized ResNeXt networks are locally in formulation similar to batch normalized recurrent neural networks, and both of them require a careful selection of initialization values for batch normalization parameters to avoid the aforementioned difficulties.

For comparison between layouts, we find **PreActBN** with shaking at  $\gamma_0 = 0.05$  also outperforms most of the settings in **PostAct** and **RPreAct** layouts with statistical significance, except for the reduction on the generalization gap by **PostAct** at  $\gamma_0 = 0.2$ , where **PreActBN@0.05** is not better with significance. We summarize the complete comparison results in Table 9.

Architecture Pair	$\gamma_0$	Valid UA	Gap
PreActBN@0.05 vs PostAct	1.00	$2.37 \times 10^{-3}$	$4.12 \times 10^{-2}$
	0.20	$2.96 \times 10^{-3}$	$6.44 \times 10^{-2}$
	0.10	$1.71 \times 10^{-2}$	$2.15 \times 10^{-2}$
	0.05	$2.21 \times 10^{-2}$	$1.27 \times 10^{-2}$
PreActBN@0.05 vs RPreAct	1.00	$1.29 \times 10^{-2}$	$1.00 \times 10^{-2}$
	0.20	$2.85 \times 10^{-2}$	$2.44 \times 10^{-3}$
	0.10	$5.07 \times 10^{-4}$	$1.34 \times 10^{-2}$
	0.05	$4.14 \times 10^{-3}$	$4.23 \times 10^{-2}$

TABLE 9: P-values from one-sided paired t-test between **PreActBN** at  $\gamma_0 = 0.05$  and **PostAct** and **RPreAct** with various value of  $\gamma_0$

Finally, when comparing  $\gamma_0$ -initialized batch normalized networks (when  $\gamma_0 = 0.05, 0.1, 0.2$ ) with their counterpart by unit initialization, we find that although both **RPreAct** and **PreActBN** show a tendency to reduce the Gap with

Architecture	Shake	$\gamma_0$	Valid UA (%)	Gap (%)
PreAct	✓	1.00	64.973	1.791
	✓	1.00	65.544	8.034
	✓	0.20	65.679	‡4.408
PreActBN	✓	0.10	66.170	6.257
	✓	0.05	‡66.432	‡5.539

\* average over three runs

‡ significantly ( $p < 0.05$ ) outperforms the same layout at  $\gamma_0 = 1.0$

TABLE 10: Performances of **PreAct** and **PreActBN** with sub-band shaking for speech emotion recognition, where  $\gamma_0$  is the initialization value of the standard deviation parameter  $\gamma$  in batch normalization

a smaller  $\gamma_0$ , only **PreActBN** manages to outperform its unit-initialized counterpart (denoted by ‡). It may suggest that the extra batch normalization between residual blocks in **PreActBN** makes it structurally more similar to batch normalized recurrent neural networks. On the other hand, any two batch normalization layers in **PostAct** or **RPreAct** are always separated by one convolutional layer.

In Section 6.1, we have seen that sub-band shaking leads to an improvement on the UA but also enlarges the generalization gap. However, instead of **PreActBN** we employed **PreAct** previously and did not experiment with different initialization values of  $\gamma$ . To present a complete comparison between shaking on the full band and on the sub-bands independently, we conduct experiments on sub-band shaking with the **PreActBN** layout and with the chosen initialization values of  $\gamma$ . The results are summarized in Table 10. Again, comparing **PreAct** and **PreActBN** both at  $\gamma_0 = 1.00$ , we could easily observe a significant increment of the generalization gap along with the additional batch normalization. With a smaller initialization value, not only does **PreActBN** reduce the generalization gap but also improve the UA when  $\gamma_0 = 0.05$  with statistical significance. Yet, the resulting 5.539% on the Gap is still significantly larger than 1.791% on the Gap by the **PreAct** layout.

We further compare the performance of **PreActBN** with shaking on the full band (66.418%/3.416%) and on the sub-bands independently (66.432%/5.539%). The p-values for these two settings are 0.51 and 0.079 for testing between the UAs and the Gaps. In the end, shaking on the full band and on the sub-bands independently give comparable performances with shaking on the full band slightly better at reducing the generalization gap.

## 7 CONCLUSION

We investigate the recently proposed Shake-Shake regularization and its variants for classification in general and speech emotion recognition in particular. In order to explain the observed interaction between batch normalization and shaking regularization, we base our ablation analysis of batch normalization in the MNIST experiments on discriminative feature learning. Our experiments show that the batch normalization right before shaking regularization is crucial in that it keeps a dispersed symmetric distribution of intermediate representations from being tilted by random perturbation due to shaking. Without the batch normaliza-

tion, shaking regularization could easily tilt the distribution of each classes and cause them to overlap each other.

In addition, we highlight the subtle difference in the requirement of embeddings for classification tasks and verification tasks, where according to the vicinal empirical minimization principle or the recent success of Mixup [18], classification tasks should try to minimize the margin between classes in feature space, while verification tasks aim to find an embedding distribution that has a minimal intra-class variation but a maximal inter-class dispersion. From this perspective, we find the embeddings by **PreActBN** with shaking are indeed distributed with small or zero margin between classes (Fig. 4(d)), while the embeddings by **PreActBN** without shaking are distributed with large margins (Fig. 4(c)). Moreover, since the original embeddings are distributed close to the class center vectors, the distribution of testing embeddings, which is often assumed to be the same or similar distribution of training samples, becomes more compact and hence leads to a higher accuracy. This finding provides a direct explanation based on the VRM principle for the ability of shaking regularization to help improve classification tasks. Another set of experiments on CIFAR-10 further validates it is the direct concatenation of batch normalization and shaking that contributes the most to the improvement on classification accuracy, and other layers of batch normalization are auxiliary.

In addition, we find that the formulation of batch normalized residual blocks resemble batch normalized recurrent neural networks. Moreover, both of these two architectures suffer from a reported issue of fast convergence but more over-fitting. To reduce the observe increment of the generalization gap in our speech emotion recognition experiments, we properly initialize the  $\gamma$  parameter with a smaller value. The experimental results validate our hypothesis and give a significant reduction on the generalization gap while achieving the same or a higher improvement on the UA.

A final comparison between shaking on the full band and on the sub-bands independently shows that with the additional batch normalization in **PreActBN** and a proper initialization value of  $\gamma$ , the difference between these two kinds of shaking is minimized, only that shaking on the full band slightly better at reducing the generalization gap.

## REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional Sequence to Sequence Learning," 2017, arXiv:1705.03122.
- [3] C. Huang and S. S. Narayanan, "Deep Convolutional Recurrent Neural Network with Attention Mechanism for Robust Speech Emotion Recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2017.
- [4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, Jan. 2014.
- [5] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *arXiv:1611.05431*, 2016.
- [6] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

- [7] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," in *BMVC*, 2016.
- [8] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep Networks with Stochastic Depth," in *ECCV*, 2016.
- [9] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the Disharmony Between Dropout and Batch Normalization by Variance Shift," *arXiv:1801.05134*, 2018.
- [10] X. Gastaldi, "Shake-Shake Regularization," in *International Conference on Learning Representations Workshop*, 2017.
- [11] Y. Yamada, M. Iwamura, and K. Kise, "ShakeDrop Regularization," *arXiv:1802.02375*, 2018.
- [12] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *Technical Report*, 2009.
- [13] N. X. Vinh, S. M. Erfani, S. Paisitkriangkrai, J. Bailey, C. Leckie, and K. Ramamohanarao, "Training Robust Models Using Random Projection," in *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, 2016.
- [14] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal Risk Minimization," in *Advances in Neural Information Processing Systems*, 2000.
- [15] T. DeVries and G. W. Taylor, "Dataset Augmentation in Feature Space," in *International Conference on Learning Representations Workshop*, 2017.
- [16] N. V. Chawla, K. W. Bowyer, L. O. Hall, , and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, 2012.
- [17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, , and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *International Conference on Learning Representations*, 2018.
- [19] V. Verma, A. Lamb, C. Beckham, A. Courville, I. Mitliagkas, and Y. Bengio, "Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer," *arXiv:1806.05236*, 2018.
- [20] Y. LeCun and C. Cortes, "MNIST Handwritten Digit Database," 2010.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016.
- [22] C. Huang and S. S. Narayanan, "Shaking Acoustic Spectral Subbands Can Better Regularize Learning in Affective Computing," in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, 2018.
- [23] C.-W. Huang and S. S. Narayanan, "Stochastic Shake-Shake Regularization for Affective Learning from Speech," in *Proceedings of Interspeech*, 2018.
- [24] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-Margin Softmax Loss for Convolutional Neural Networks," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 507–516.
- [25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [26] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [27] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *arXiv:1801.07698*, 2018.
- [28] X. Qi and L. Zhang, "Face recognition via centralized coordinate learning," *arXiv:1801.05678*, 2018.
- [29] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning An Invariant Mapping," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
- [30] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A Unified Embedding for Face Recognition and Clustering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [31] V. Vapnik, *Statistical Learning Theory*. J. Wiley, 1998.
- [32] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion Recognition Based on Phoneme Classes," in *Proceedings of InterSpeech*, 2004.
- [33] H. Mallidi and H. Hermansky, "Novel Neural Network Based Fusion for Multistream ASR," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2016.
- [34] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift)," *arXiv:1805.11604*, 2018.
- [35] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," in *International Conference on Learning Representations*, 2014.
- [36] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch Normalized Recurrent Neural Networks," in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, 2016.
- [37] T. Cooijmans, N. Ballas, C. Laurent, Ç. Gülcöhre, and A. Courville, "Recurrent Batch Normalization," in *International Conference on Learning Representations*, 2017.
- [38] O. Martin, I. Kotsia, B. M. Macq, and I. Pitas, "The eINTERFACE'05 Audio-Visual Emotion Database," in *Proceedings of the International Conference on Data Engineering Workshops*, 2006.
- [39] S. R. Livingstone, K. Peck, and F. A. Russo, "RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song," in *Proceedings of the 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*, 2012.
- [40] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, Nov 2008.
- [41] F. Burkhardt, A. Paeschke, M. Rolfs, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proceedings of Interspeech*, 2005.
- [42] G. Costantini, I. Iaderola, andrea Paolini, and M. Todisco, "EMOVO Corpus: an Italian Emotional Speech Database," in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014.
- [43] S.-U. Haq and P. J. Jackson, *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, ch. Multimodal Emotion Recognition.
- [44] G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, "Convolutional Neural Networks with Data Augmentation for Classifying Speakers Native Language," in *Proceedings of Interspeech*, 2016.
- [45] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The KALDI Speech Recognition Toolkit," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.
- [46] "Pytorch: Tensors and Dynamic Neural Networks in Python with Strong GPU Acceleration," 2017, <https://github.com/pytorch/pytorch>.
- [47] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, 2014.