Noise-based and class-based curriculum learning for image classifiers

1st Yue Li
Riga Technical University
Riga, Latvia
yue.li@leesensei.com

2nd Evalds Urtans
Riga Technical University
Riga, Latvia
evalds.urtans@rtu.lv

***Abstract*—Datasets often contain different difficulty samples and even noisy samples. This paper introduces two naive curriculum learning methods, one using an image dataset with noise and another one using an image dataset that contains samples from other datasets with presumed higher difficulty. The final goal is to improve the performance of the model by gradually introducing more difficult samples during the training process rather than using them from the very beginning. Experiments demonstrated that using the proposed curriculum learning methods, a classifier can achieve higher accuracy in less training epochs.**

***Keywords*—Deep Learning, Curriculum Learning, Image Classification**

I. Introduction

Teaching natural intelligence by gradually increasing the difficulty of learning materials has been used for a long time.

This idea has been introduced to the field of machine learning, and the term "curriculum learning" was coined in the paper [1].

At its core, curriculum learning is the design of a training strategy that often comes down to the order of samples that are fed into a model.

This is similar to the way courses are taught in educational institutions.

For example, students are taught the concept of Gaussian distribution first and only after that the teacher will introduce Central Limit Theorem. This paper introduces two naive curriculum learning methods for training image classifiers.

[Noise-based curriculum learning] Using a clean version of the original dataset and starting from $p = 0\%$ noise, noise-based curriculum learning increases the percentage of noise in the dataset at each epoch during training by $\frac{p_{max}}{epochs}$, where p_{max} is the maximum percentage of noise that will be reached by the end of the training.

[Class-based curriculum learning] Using a clean dataset and starting from a single included class of a more complicated dataset, class-based curriculum learning increases the number of classes in the data set by one class at each of the predefined steps during training until the maximum number of classes has been reached. A mixed dataset contains classes of higher difficulty at the end of the training.

II. Related work

Curriculum learning has been used in different settings since its related to training strategies in general. Several articles discuss the use of naive curriculum learning methods. In some research, naive curriculum learning has been used in the medical domain by using smaller image patches to train a classifier first and then feed complete images to the model [2]. Other research has used multistep curriculum learning to obtain better performance in the localization of thoracic disease [3]. And yet other research has used curriculum learning for the classification of visual attributes [4].

III. Methodology

The purpose of this paper is to demonstrate the effectiveness of learning strategies for noise-based and class-based curriculum learning methods.

A controlled experiment was used. Each experiment has one control group (curriculum learning group) and one experimental group (non-curriculum learning group), and the only difference between these groups was the curriculum learning strategy.

Each of the experiments and sub-experiments was repeated 10 times to gain statistical significance.

A. Datasets

Four datasets were used to evaluate the proposed curriculum learning methods. MNIST, KMNIST and CIFAR10 datsets were used for evaluating the noise-based curriculum learning. KEMNIST dataset, a mixture dataset consisting of images from the EMNIST letter dataset and the KMNSIT dataset, was used to evaluate the class-based curriculum learning. EMNIST letter dataset contains 26 classes and the KMNIST dataset contains 10 classes, which makes the total number of classes in the mixed dataset 36.

B. Metrics

The point estimator $D^i = A_c^i - A_{nc}^i$ is used to measure improvements in generalization ability when using different curriculum learning methods. In equation A_c^i denotes the test accuracy of the curriculum group and A_{nc}^i denotes the test accuracy of the non-curriculum group when the dataset is applied i percent of pepper noise (noise-based curriculum learning) or when the number of KMNIST classes that the model is allowed to sample is i (class-based curriculum learning). For noise-based curriculum learning, we will report $E(D)$ and its 95% confidence interval when the distribution

of the percent of pepper noise is distributed according to a bounded normal centered at 20. For class-based curriculum learning, we will report $E(D^i)$ and its 95% confidence interval when the maximum number of KMNIST classes that the model is allowed to sample i is 5 and 10.

IV. Experiments

A. Noise-based curriculum learning experiment

Noise-based curriculum learning has been evaluated using MNIST, KMNIST and CIFAR10 datasets. Those datasets do not contain obvious noise. Noise has been artificially added using pepper noise, as shown in Algorithm 1.

For the CIFAR10 dataset, pepper noise was applied in all three color channels.

The training set sizes of three sub-experiments are 50000, 50000 and 40000 respectively. The validation set size and test set sizes are 10000 for all three sub-experiments.

The noise curriculum is illustrated in Figure 2 on page 3.

Algorithm 1 Artificial Pepper Noise

$n \leftarrow \lfloor p * M * N \rfloor$
$\overline{I} \leftarrow Uniform(range(0, M * N), n)$
while $t < l$ **do**
  $i \leftarrow \frac{I[t]}{M}$
  $j \leftarrow I[t] \% M$
  if $i + j$ is even **then**
    $A_{ij} \leftarrow 0$
  else
    $A_{ij} \leftarrow 255$
  $t \leftarrow t + 1$

The architecture used in the noise-based curriculum learning MNIST sub-experiments is a ResNet [5]. The first resblock contains two 2D convolutional layers, the first convolutional layer has 1 input channel, 2 output channels, the second one has 2 input channels and 4 output channels; for the second resblock, there are also 2 convolutional layers, the first one has 4 input channels and 4 output channels, the second one has 4 input channels and 4 output channels, after these two resblocks, there is an additional convolutional layer with 4 input channels and 8 output channels, and finally there are two fully connected layers with 6272 and 10 channels, respectively. Adam[6] optimizer has been used. The model was trained for 20 epochs with a learning rate of 0.003 and batch size of 16.

For the KMNIST subexperiments, ResNet model with two resblocks, each containing two convolutional layers. After 2 resblocks, two pairs of convolutional layer and maxpooling layer pairs follow. SGD optimized has been used with momentum 0.9. The number of epochs, learning rate, batch size are set to 20, 0.003 and 16 respectively.

In the CIFAR10 sub-experiments, batch normalization[7] and maxpooling functions have also been added. The architecture consisted of 9 convolutional layers with in channels and out channels being the Fibonacci numbers, after every 3 convolutional layers, 1 batch normalization layer and 1 maxpooling layer are applied sequentially. SGD optimized has been used with momentum 0.9. Hyper parameters number of epochs, learning rate, batch size are 20, 0.075 and 30 respectively.

All three models used the same convolution with the $3x3$ kernel and a stride of 1, and ReLU as activation function. All validation sets and test sets were clean images with no added noise.

B. Class-based curriculum learning experiment

Two sub-experiments are designed in this experiment, requiring the same model classifying images from the EMNIST dataset containing respectively 5 and 10 classes from the KMNIST dataset. The dataset configuration contained 100000 samples in the training set, 10000 samples in the validation set, and 10000 samples in the test set. The number of the KMNIST classes is held constant for the non-curriculum group under each sub-experiment, whereas the number of KMNIST classes is changing for the curriculum group according to class curriculum. Moreover, the percentage of KMNIST samples is the same across training set, validation set, and test set for the non-curriculum group of a particular sub-experiment. For the curriculum group, the percentage of KMNIST samples in the validation set and test set is the same, but the percentage of KMNIST samples in the training set will gradually increase to the same value as in the validation set and test set.

The schematic representation of the class-based curriculum learning is given in Figure 3 on page 3.

The architecture used in class-based curriculum learning contains nine 2D convolutional layers where in channels and the out channels follow Fibonacci numbers as shown in Figure 4 on page 4. All convolution layers are the same convolution layers with a stride of 1. After every 3 convolutional layers, 1 batch normalization layer and 1 maxpooling layer are applied sequentially. ReLU is used as a non-linearity function. The learning rate is set to 0.02, batch size used is 16 and the number of epochs trained is 100.

In the class-based curriculum learning group, the addition of KMNIST classes is not done by changing the architecture.

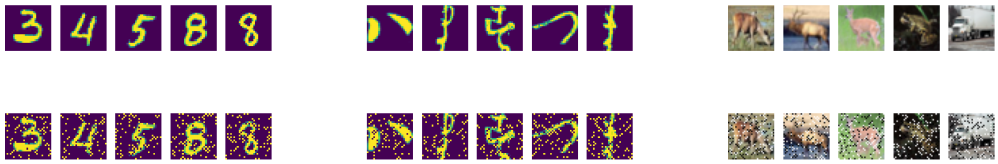
Additional KMNIST classes are added during training, allowing the model to sample more classes from the KMNIST dataset.

V. Results

In this section, we show the results obtained from emperical experiments. Some aggregation plots are given below, although many more experiments have been done to find the best hyperparameters and configurations for each curriculum and non-curriculum learning method. Only results from the same set of sub-experiments are compared.

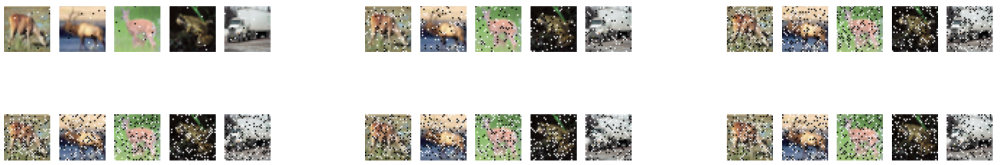
A. Noise-based curriculum learning

In Figure 5 on page 4 the test loss after 100 epochs of the models in the noise-based curriculum learning group in three sub-experiments is significantly smoother than the non-curriculum group when compared to their counterparts in the



(a) MNIST (b) KMNIST (c) CIFAR10

Fig. 1: Clean samples drawn from MNIST, KMNIST, and CIFAR10 dataset and also samples with 20% pepper noise applied.



epoch 0
curriculum group: 1% noise
non-curriculum group: 20% noise
percent

epoch 9:
curriculum group: 10% noise
non-curriculum group: 20% noise

epoch 19
curriculum group: 20% noise
non-curriculum group: 20% noise

Fig. 2: Noise-based curriculum learning for sub-experiments adding a maximum of 20% noise. The top row is the images used in training for curriculum group, the bottom row is the same images with different level of pepper noise used in training for non-curriculum group. The noise in the images for the curriculum group is gradually increased until it reaches 20%, whereas the noise on the same image for the non-curriculum is fixed at 20%.

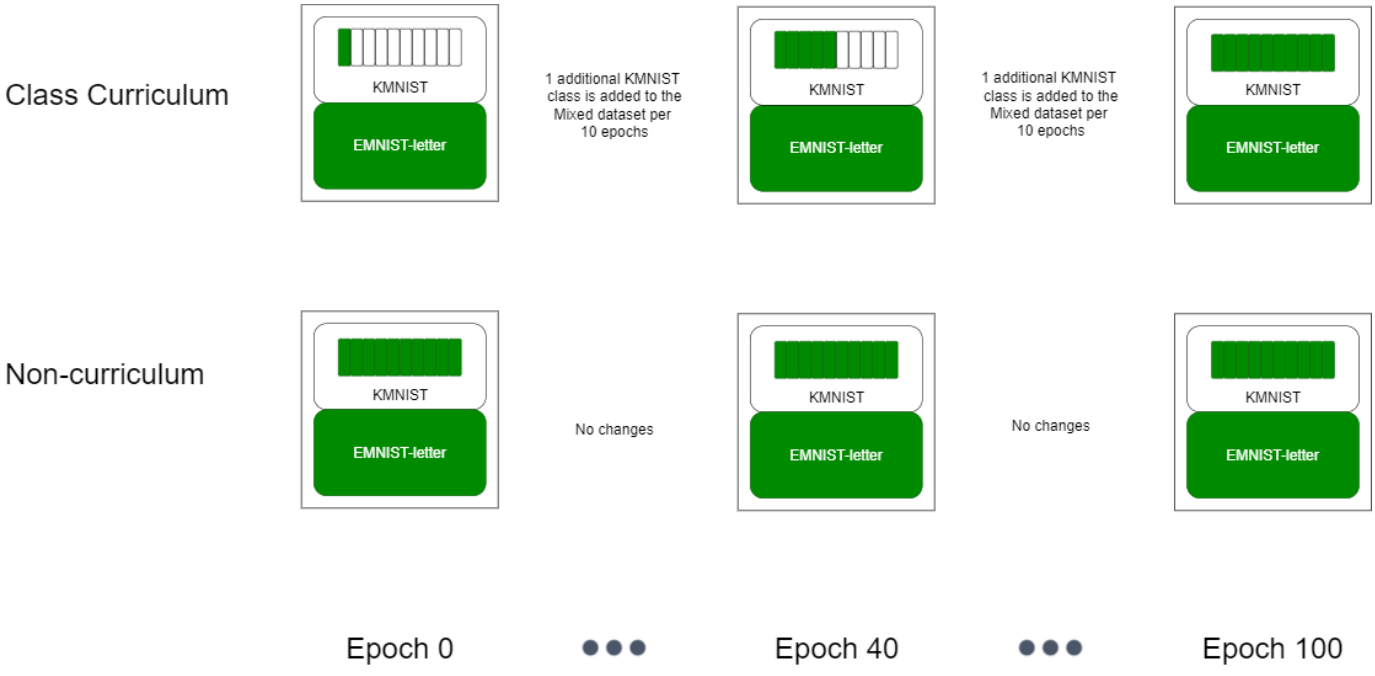


Fig. 3: Process of class-based curriculum learning. At predefined epochs new classes from more complex dataset have been aded to the training and test datasets.

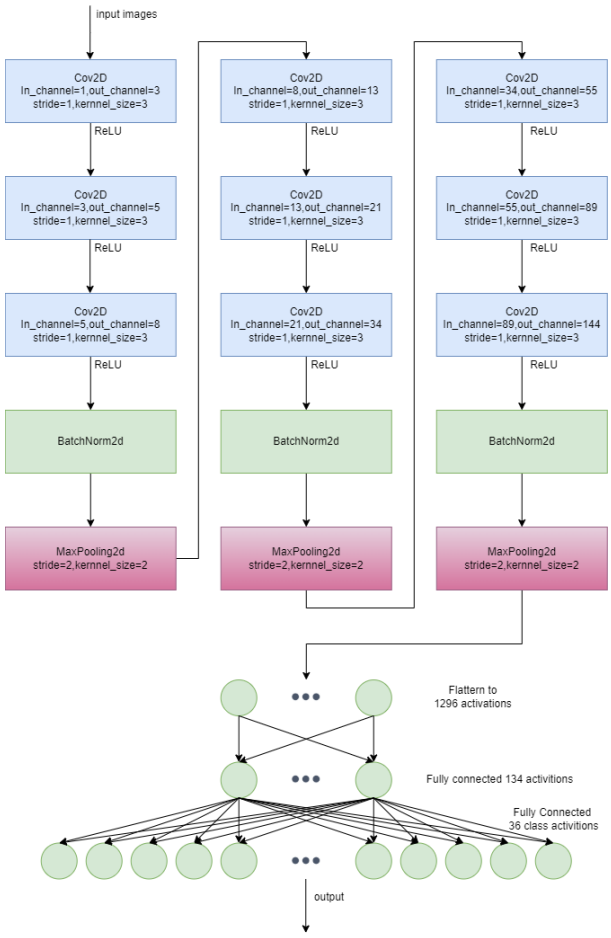


Fig. 4: Class curriculum architecture

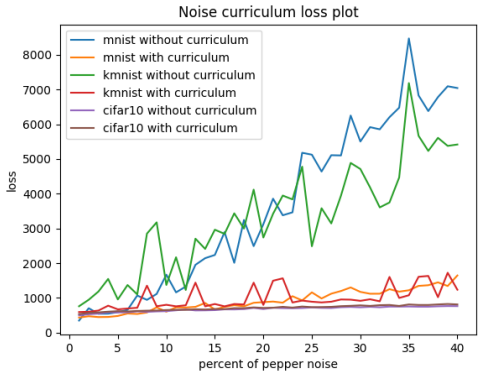


Fig. 5: Correlation between loss and percentage of noise in noise-based curriculum and non-curriculum learning methods

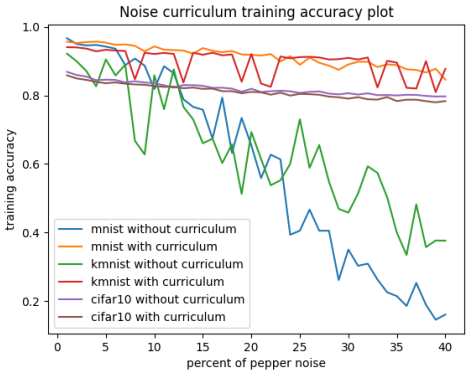


Fig. 6: Correlation between accuracy and percentage of noise in noise-based curriculum and non-curriculum learning methods

respective sub-experiments. However, when the percent of pepper noise is low, the models in the non-curriculum groups at few points perform better than the curriculum groups. For example, the blue curve intersected with the orange curve on the far left of the figure. This is because stochasticity present in the optimization algorithm gives the model in the non-curriculum group chances to surpass the curriculum group when the level of applied noise is low. These results show that noise-based curriculum learning achieves better results.

Noise-based curriculum learning experiments achieve higher accuracies than non-curriculum learning counterparts, as shown in Figure 6 on page 4. The test accuracies of the curriculum groups in the MNIST and KMNIST sub-experiments are significantly better than in their respective non-curriculum groups. The perforemance of both groups in the CIFAR10 subexperiment is very close. It could be that the reason behind this is that CIFAR10 is a difficult or noisy dataset itself, even without added noise.

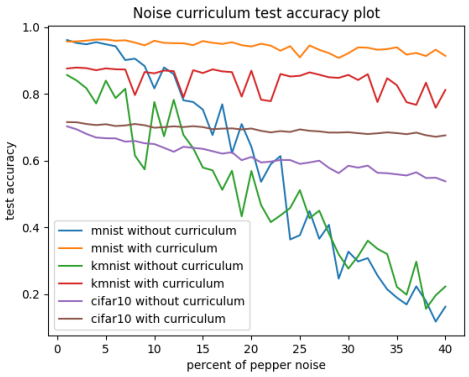


Fig. 7: Noise curriculum test accuracy plot for three sub-experiments

Although the number of pixels that are corrupted in an image takes a discrete integer value, there are a large number of pixels in an image. Therefore, it is reasonable to consider the percentage of noise to have a real value; hence, it is reasonable to think that it is distributed according to a PDF.

Because the prior and posterior densities of the percentage of pepper noise appearing in an image are unknown, it has been assumed that the percentage of pepper noise appearing in the training set follows a bounded Gaussian distribution in the range $[0, 40]$ and centered at 20 with variance of 1. We used the PMF of Bin $\sim (40, 1/2)$ to approximate the normal and calculated the mean, standard deviation, and confidence interval of the random variable D for three datasets. The numerical results are shown in I. All the bounds of the three confidence intervals are greater than 0. With the assumptions held and under the experiments settings, noise curriculum can improve generalization performance on these three datasets with approximately 95% confidence.

TABLE I: Noise-based curriculum learning results

Dataset	Mean	Std	C.I.
MNIST	0.3297	0.1631	[0.213,0.4463]
KMNIST	0.3366	0.0873	[0.2741,0.3991]
CIFAR10	0.0836	0.0241	[0.0663,0.1009]

B. Class-based curriculum learning

Class-based curriculum learning results show a significant improvements in accuracy compared to non-curriculum learning methods.

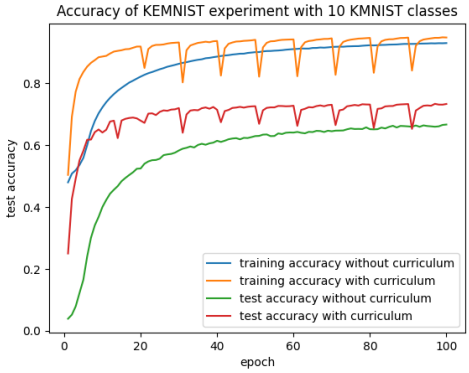


Fig. 8: Accuracy of class-based curriculum training with 10 KMNIST classes incrementaly added during the training

The loss values in the learning group of the curriculum have several peaks as shown in Figure 9 on page 5.

This is because new KMNIST classes have been added during training at the specified epochs. However, the loss values in the curriculum groups quickly drop below their respective non-curriculum groups and eventually stay below after all KMNIST classes are added.

The accuracies follow some pattern as illustrated in Figure 8 on page 5.

Two figures of training and test accuracies are also provided for two sub-experiments. Instead of peaks, the curves for the curriculum groups now have valleys. Again, for the same reason that new classes have been added at those specific epochs.

After addition, both the training accuracy and test accuracy quickly recover above their respective non-curriculum groups and eventually stay above it.

This shows that class-based curriculum can help models learn better knowledge from samples during training and help increase generalization capabilities of a model incrementally reaching higher performance.

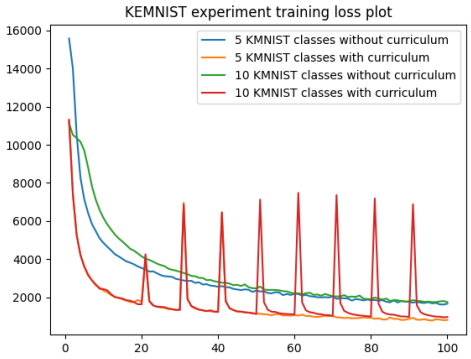


Fig. 9: Comparison of loss between class-based curriculum and non-curriculum learning methods

TABLE II: Class-based curriculum learning results

Number of KEMNIST Classes	Mean	Std	C.I.
5	0.0525	0.0428	[0.015,0.0901]
10	0.1217	0.0869	[0.0347,0.2088]

The statistics of the class-based curriculum learning experiment are computed using data acquired from 10 repeat runs and are presented in Table II. The confidence intervals for the mean of D^i from two sub-experiments with, respectively, 5 and 10 KMNIST classes both do not reach a value of 0. This means that our class curriculum is able to increase the model's performance under the settings of this experiment with 95% confidence.

VI. Limitations

There are multiple limitations on the results obtained using curriculum learning methods described in this document.

In noise curriculum experiments, the same percentage of noise was applied to all training samples at each stage of training. Peper noise that has been added using a bounded Gaussian distribution in the interval [0,40] and centered at 20, because noisier images usually becomes unrecognizable.

For the learning method of the class curriculum, the results look promising, but the number of sub-experiments could be too low.

VII. Further research

This paper examined two naive curriculum learning methods that the authors have manually configured using academic datasets to illustrate and test the proposed methods. Designing and tuning curriculum learning methods is time consuming.

Future research cloud should shift focus to automated curriculum learning method setup. It should be possible to estimate the noisiness or complexity of the samples and then apply the methods proposed in this paper. Also, more complex, larger, and real-world datasets should be studied.

VIII. Conclusions

This article has shown that if there is prior knowledge of the cleanliness of a data set or its complexity of classes, then it can be incorporated into the training process by using noise-based or class-based curriculum methods. This paper used scientifically rigorous controlled experiments to examine these approaches. Findings of this research show statistically significant improvement in performance of a model when testing it using clean images from MNIST, KMNIST and CIFAR10 datasets while training with noise-based curriculum learning method. The results also show that the class-based curriculum learning method can achieve better generalization performance on a mixed test dataset comprising all of the classes from both datasets.

Acknowledgements

Research has been completed with support from the High Performance Computing Center of Riga Technical University, which provided 12 nVidia K40 GPUs and 8 nVidia V100 GPUs.

References

[1] Y. Bengio, J. Louradour, R. Collobert, , J. Weston, "Curriculum learning," in *ICML '09*, 2009.

[2] A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. Á. G. Ballester, , G. Piella, "Medical-based deep curriculum learning for improved fracture classification," in *MICCAI*, 2019.

[3] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, , R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *MLMI@MICCAI*, 2018.

[4] N. Sarafianos, T. Giannakopoulos, C. Nikou, , I. Kakadiaris, "Curriculum learning of visual attribute clusters for multi-task classification," *ArXiv*, vol. abs/1709.06664, 2018.

[5] K. He, X. Zhang, S. Ren, , J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[6] D. P. Kingma , J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[7] S. Ioffe , C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ArXiv*, vol. abs/1502.03167, 2015.

[8] Y. A. LeCun, L. Bottou, Y. Bengio, , P. Haffner, "Gradient-based learning applied to document recognition," 1998.

[9] S. Zhang, X. Zhang, W. Zhang, , A. Søgaard, "Worst-case-aware curriculum learning for zero and few shot transfer," *ArXiv*, vol. abs/2009.11138, 2020.

[10] M. G. i Calabuig, C. Ventura, , X. G. i Nieto, "Curriculum learning for recurrent video object segmentation," *ArXiv*, vol. abs/2008.06698, 2020.

[11] Y. Wang, W. Gan, W. Wu, , J. Yan, "Dynamic curriculum learning for imbalanced data classification," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5016–5025, 2019.

[12] D. Weinshall , G. Cohen, "Curriculum learning by transfer learning: Theory and experiments with deep networks," *ArXiv*, vol. abs/1802.03796, 2018.

[13] J. Wang, X. Wang, , W. Liu, "Weakly- and semi-supervised faster r-cnn with curriculum learning," *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2416–2421, 2018.

[14] W. Lotter, G. Sorensen, , D. D. Cox, "A multi-scale cnn and curriculum learning strategy for mammogram classification," in *DLMIA/ML-CDS@MICCAI*, 2017.

[15] X. Liu, P. He, W. Chen, , J. Gao, "Multi-task deep neural networks for natural language understanding," in *ACL*, 2019.

[16] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, , W. chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015.

[17] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[18] X. Glorot, A. Bordes, , Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, 2011.

[19] Y. LeCun , C. Cortes, "The mnist database of handwritten digits," 2005.

[20] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[21] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, , D. Ha, "Deep learning for classical japanese literature," *ArXiv*, vol. abs/1812.01718, 2018.