Microsoft Word Author Guidelines for CVPR Proceedings

Explicit Curriculum Learning

Abstract

Image blur and additive image noise are common examples of image distortions that can emerge during the acquisition process. The most prominent techniques for increasing the resilience of DNN models to these and other types of distortions are based on the training set augmentation and on the curriculum learning ideas. We suggest a model-based approach for combining these two methods into a single training framework that we call Explicit Curriculum Learning (ECL).

In ECL, the expected test-time image distortions are explicitly modeled using discrete or continuous distributions over the distortion parameters (e.g. over blur window size or/and over AWGN power). The training curriculum is constructed using a sequence of distributions gradually converging to the target test-time distribution. The sequence starts from producing mostly pristine images and gradually introduces more distorted and, thus, more difficult examples as training is progressing. We also suggest a general method for constructing training curriculum using mixture distributions.

We evaluate the suggested ECL method using CIFAR-10 benchmark and a state-of-the-art CNN-based network. The presented results demonstrate that ECL approach outperforms the conventional training method and helps to produce more robust DNN models without altering underlying network architecture.

1. Introduction

In recent years there was a great progress in many image classification problems due to advances in the development of deep neural networks [1,2,3]. Networks are usually trained with large quantities of training examples. The success of the training procedure relies on the assumption that the properties of the training data are similar to the properties of the data encountered during the inference. For instance, if a certain distribution of noise or distortion effects is expected to be present in the test data, the training data should have the same distribution of the noise and distortion parameters. Therefore, clean (pristine) training data is frequently augmented by adding synthetic noise and artificial distortions (e.g. rotations) for training robust deep learning models. Generally, the augmentation of input images can be considered as a form of regularization. Notably, Res-Net [18] and VGG-Net [8] architectures achieved a significant improvement in image classification and image recognition challenge [25] with various data augmentation techniques such as padding and horizontal flipping.

However, adding high level of noise and strong distortion effect makes training difficult since it might blur the boundaries between different classes. As a result, neural networks trained with augmented data might have low accuracy on pristine examples [3,4,5,6]. Therefore, there is a trade-off between the level of robustness of the model and its accuracy on the pristine inputs. The conflict is especially aggravated when the model is required to handle infrequent noisy images while being stable on pristine images that occurs most of the time.

The robustness of DNN model can be improved using the curriculum learning approach [7]. In this method, a robust network is trained by presenting training examples according to a certain schedule that starts from “simple” training examples and gradually introduces more “difficult” examples as training is progressing. For instance, an example is considered difficult if it falls on the incorrect side of the decision surface of the Bayes classifier.

In this work we suggest the Explicit Curriculum Learning (ECL) framework for training robust deep neural network models. The ECL method combines the curriculum learning approach with a controlled augmentation of the training set based on explicit modeling of the target distribution of the distortion parameters.

In ECL, the curriculum schedule is defined by a sequence of distributions that are used for augmenting training examples as the training is progressing. The idea is to build the sequence in a way that ensures that more difficult (i.e. more distorted) examples are presented with each next training epoch. With each training epoch, the distribution of the training batch is made a bit closer to the assumed noise distribution within the test data. In this way, we start training the network with simple, clearly separable examples. Once the network learns to recognize simple examples, it’s presented with a training set that includes a larger proportion of more difficult noisy examples for expanding and refining class boundaries.

We describe a general recipe for constructing training curriculum using a sequence of mixture distribution. The training set at each curriculum step is built by mixing the target distribution of distortion parameters with a parametrized relaxed distribution that generates “easy” training examples.

We demonstrate the validity of this approach using the task of image classification (using CIFAR-10 data set [14]). We evaluate the performance of the suggested training procedure and compare it to the conventional training. We demonstrate that the proposed training procedure results in the performance improvements without any change to the underlying network architecture.

The rest of the paper is organized as following: section 2 reviews previous work with emphasis on curriculum learning approach that forms a basis for our method that is summarized in section 3. In section 4, we describe the data generation process and evaluation methodology that is used through the following section. In section 5 we describe a simple toy example for explicit construction of curriculum schedules. We extend this simple example by considering continuous models and a mixture-based recipe for constructing curriculum in section 6. Section 7 provide a summary of the proposed method and describe some possible future directions

2. Previous Work

Deep Neural Networks is widely used in computer vision applications. The successful application of DNN is fuelled by the availability of various large datasets for training and evaluation. Using this dataset, DNN and CNN in particular allowed to achieve almost human performance in many computer vision tasks [9]. There also have been many efforts for achieving a greater degree of robustness to noise and other image distortions with DNN models [13,24].

Various techniques to deal with distortions relies on applying restoration methods prior to classification. The authors in [21] presented a novel technique of combining multiple stacked sparse denoising autoencoders for image classification. In [22], CNN was used for removing multiple types of distortions for image recognition. An interesting approach for learning to deblur with convolutional neural networks is proposed by [23]. In this approach a learning-based deep structure for blind image deconvolution is employed. Although achieving very good results for smaller kernels, their method is only for image deconvolution without considering further classification step.

Data augmentation is often used to reduce overfitting on models, where we increase the amount of training data using information only in our training data. For instance, on ImageNet, the data augmentation approach by [2], introduced in 2012, remains the standard with small changes. Although data augmentation improvements have been found for a particular dataset, they often do not transfer to other datasets as effectively. For example, horizontal flipping of images during training is an effective data augmentation method on CIFAR-10[14], but not on MNIST [13], due to the different symmetries present in these datasets. The augmentation of images using additive noise or applying random transformations is a common practice for training robust noise-resilient deep learning models [3,4,5,17].

In recent years, the development of noise-robust neural networks has been investigated. For instance, in 2014 [24] presented deep hybrid networks that can cope with some types of noisy images in image recognition, while [13] designed a network that can deal with noise in speech recognition. However, publicly available state-of-the-art networks such as VGG-net, Res-Net and RNN’s, which considered as standards, are yet to achieve impressive results on distorted data, which indicates the heavy effect of noise on performance.

Our approach is based on the idea of curriculum learning [7]. In this approach, the network is presented initially with “easy examples” (e.g. undistorted images) while gradually adding more distorted images to the training mix.

In the original settings, the target distribution of the inputs ‘q’ is assumed to be P(q). The training examples are drawn at each step (e.g. epoch) according to a parameterized sequence of distribution $Q_\lambda(q)$ with $Q_{\lambda=1}(q) = P(q)$. The sequence gradually converges to the target distribution with intention to introduce more diversity and ,thus, more difficulty into the training set as the parameter λ, a quality indicator for each distribution of sequence, increases. Without loss of generality, we assume that λ is a real number between 0 and 1. λ =0 means a sequence of training containing only the highest quality data and when λ =1 ,the sequence data is equal to the target distribution (e.g. $P(q)$).

In the original paper, authors use the entropy [7] as a measure of difficulty, Namely, it’s requires that the entropy should increase with the parameter $H(Q_\lambda) < H(Q_{\lambda+\varepsilon})$.

The original work demonstrates the idea of the Curriculum Learning (CL) using three toy classification problems. These problems are used for introducing three different methods for building a training curriculum (a sequence of training distributions as above). The first method suggests using a synthetically generated data for training an SVM classifier for a simple binary classification problem. Since the data is artificially generated, the true decision surface is known *a priori*. Therefore, the data can be labelled according to its distance to the true classification decision surface while considering examples that fall on the wrong side of the surface as being the most difficult. Although this approach demonstrates the idea of curriculum learning, it’s not applicable in practice since the true decision surface is not known. The second example provided in the paper introduces random attributes into input vector. The number of nonzero random (irrelevant) attributes is gradually increased with each training iteration. This approach is also not practical as partition into relevant and irrelevant input features is not known apriori in the practical setting. The third example builds a shape (rectangle, ellipse, triangle) classifier from images. The classifier trained with gradually more complex training instances of each shape (i.e. occupying less space in the image or having more variability in shape parameters).

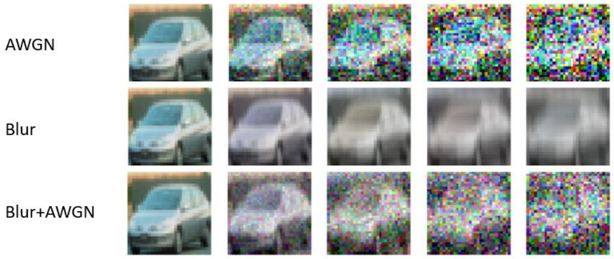

Figure 1: Examples of Cifar-10 with various distortions

In recent years, curriculum learning has gained more attention from the research community. The authors in [28], presented a curriculum-based CNN for scene classification. The described training curriculum is based on image difficulty that is defined by the source of an image. (Google vs. Flickr). In curriculum-based image segmentation,[29] an initial segmentation model is trained with simple images using saliency maps for supervision. Then, the samples of increasing complexity are progressively included to further enhance the ability of the segmentation model. In [30], the authors investigated the robustness of curriculum learning in common computer vision tasks and highlighted the superiority in convergence.

3. Our contribution

Contrary to the original curriculum learning approach we start from an explicit probabilistic model of the target noise distribution and provide a general recipe for building a curriculum by mixing the target distribution with a relaxed distribution(Beta distribution in our experiments) for controlling the difficulty of the training set through a number of parameters.

In this work we have followed the above approach by considering a standard deep learning task (CIFAR10 [14]) and by training CNN with reported state-of-the-art results for this task [1,3]. We have evaluated the robustness of the conventional training models using augmented training examples. We compare these baseline performance evaluations with the performance obtained using various curriculum models (e.g. various parameterized families of curriculum distributions). We conjecture that the curriculum approach will result in more robust models and, perhaps, will allow to reduce the number of iterations for training the robust neural networks.

4. Data and evaluation framework

This section describes the experiments conducted to assess the performance of image classification task using the proposed curriculum learning schemes. First, we describe the CIFAR-10 [14] dataset, network architecture and data augmentation methods used in the experiments. Then, we present two different cases for approaching the target noise distribution, namely discrete schedule and continuous schedule.

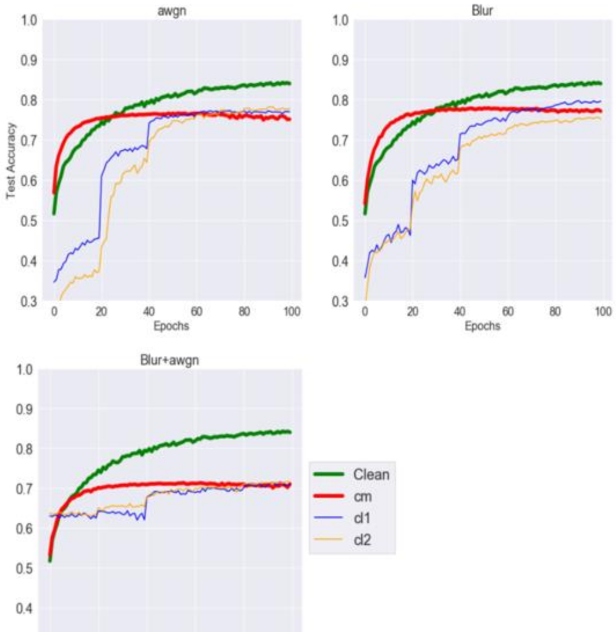

Figure 2: Test accuracy with respect to epochs – explicit discrete model. Comparison of the curriculum strategies to baseline for all three settings

Dataset. We conduct our experiments on a well-known dataset named CIFAR-10 [14]. CIFAR-10 consists of 60,000 32 × 32 colour images in 10 classes, with 6,000 images per class. 50,000 are used for training, and 10,000 are test images.

Network architecture. For all of the following experiments we used CNN from Keras open source library. All hyperparameters remain the same for all experiments with 100 epochs, batch size of 128 and RMSprop as optimizer.

Data Augmentation Let $\{(X_i, Y_i\}$ denote a classification data set, where i-th data point consist of input x_i (an image) and y_i its corresponding label. We define q_i as a quality indicator for each data point. Without loss of generality, q_i is a real number between 0 and 1 where $q_i = 0$ is the highest quality (pristine image) and $q_i = 1$ is the lowest quality (i.e. highest noise level). Although the quality indicator of individual test example is unknown, we control the quality of training example by varying q_i which we call the quality mix pdf or $QPDF_\lambda(q)$. In the conventional training methods, the training set has the same QPDF as the expected test data. In our curriculum learning approach, we construct a sequence of quality mix distributions ($QPDF_\lambda(q)$) that approach the target (test) QPDF gradually as training is progressing. At the end of training (that is, in the last epochs) the QPDF is equal to the target QPDF.

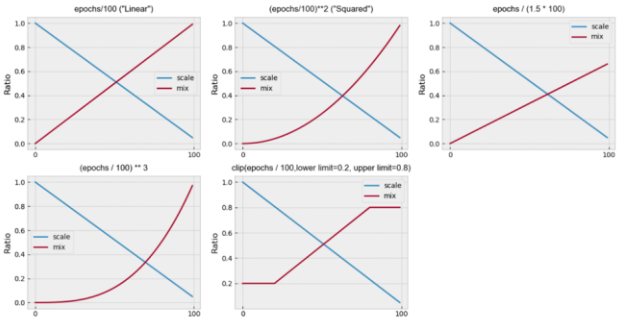

Figure 3: Different schedules for curriculum learning, based on curriculum function with custom normalization for proper visualization. Each graph represents the data mixture as training progresses. 'Scale' refers to the beta distribution, while 'mix' refers to lambda parameter. Our experiment showed that both linear and squared mix distributions got improved results with decrease in generalization error compared to baseline, while other curriculums perform worse. Thus, suggesting that a curriculum learning approach needs fine tuning in order to generalize better.

We conduct our experiment in the presence of two types of noise, additive white gaussian noise (AWGN) and blur. For AWGN case, the highest level of noise (i.e. $q_i = 1$) is defined with the standard deviation of 80. Similarly, the highest distortion level of blur is with a 7x7 kernel size. Examples of image distortions are shown in *Figure 1*. The baseline approach does not consider any curriculum and contains all the data set. The training set has the same QPDF as test. We term the baseline as conventional method (CM). As an upper bound baseline, we consider using for train and test only pristine images. All other settings, such as network architecture and hyperparameters are identical for all experiments.

5. Explicit Discrete Model

We start by considering a simple explicit discrete model for describing the target distribution of distortion parameters. We use a simple, stepwise schedule for approaching the target distribution and evaluate the resilience of the resulting model.

In the case of discrete schedule, we use a fixed size of noise levels, i.e. Q is a step function growing as training is progressing. In fact, q_i is proportional to the standard deviations in AWGN and similarly to kernel size in blur.

We considered three settings for which we compare our curriculum learning approach to CM. **Settings 1**: noise levels contain only AWGN with varying standard deviations of {20,40,60,80}.**Settings 2:** noise levels are increasing blur impact with different kernel size of {2x2,3x3,5x5,7x7}. **Settings 3:** noise level represents a mix of blur and AWGN. First, we apply blur with the same kernel sizes as in settings 2. Then, we add AWGN to blurred images with the same standard deviations as in settings 1. This setting will generate the most corrupted data. Please note that distorted data is added to the original pristine images, thus containing five levels of noise in all three settings.

In order to create a proper curriculum, we constructed two approaches to reach target distribution. The first method introduces the network with an entire set of next noise level in increasing order every 20 epochs. In the second method, we introduce the network with half quantity of the set with next noise level in increasing order, every 10 epochs. Since these methods suggest that training set quantity will be higher in the conventional method, we generate additional noisy data with same noise level for our curriculum-based approach. For both curriculum methods, we constructed a schedule that contains an equal proportion of noise levels every 20 epochs until convergence to target QPDF is reached.

For all three settings we compare our two curriculum learning strategies with the conventional method and clean data method. From *Figure 2,* we can see that all models are sensitive to the presence of noise as the classification performance decreases dramatically compared to the upper bound. This is in accordance with the evaluation results in [1,20]. This degradation can be explained by the fact that distortions can heavily remove the texture and edge information in an image, since DNN models may always attempt to look for specific textures and edges for the classification task. In addition, our first curriculum-based strategy outperforms the conventional method in first two settings, with 33% and 22% improvement in accuracy, respectively. In the case of adding blur and AWGN as image distortions (settings 3), the first curriculum strategy does not improve results, while the second method increase test accuracy by 10% compared to baseline. Interestingly, in our experiments, the conventional method converges much faster than both curriculum strategies which contradicts the notion of faster convergence with curriculum based approaches, as stated in [7] .

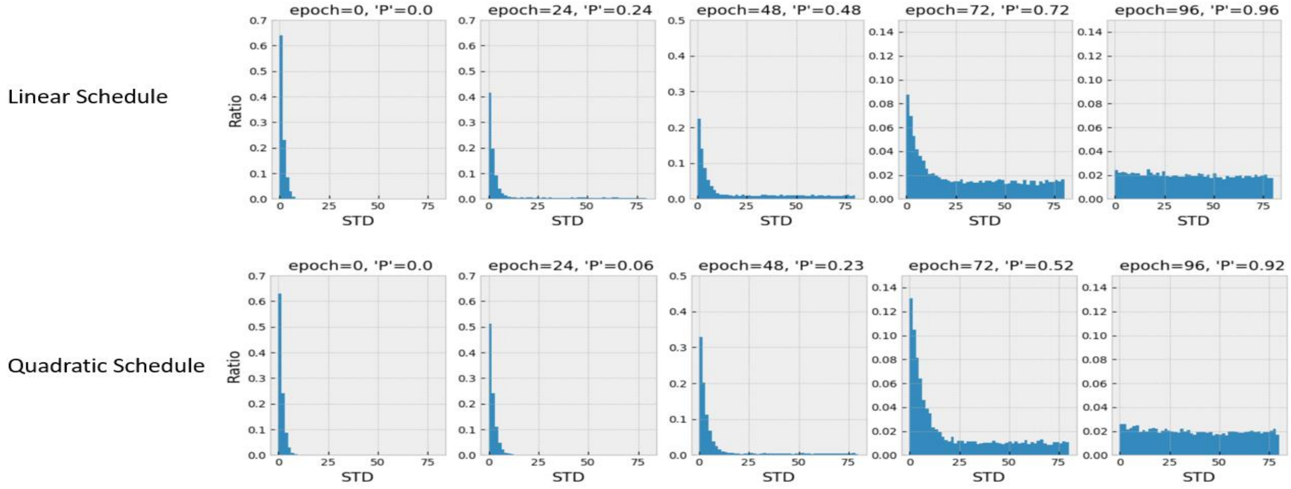

Figure 4: Noise levels of uniform target distribution with respect to AWGN standard deviation for linear and quadratic schedules. Notice the slightly faster convergence to target distribution in linear schedule.(for other curriculum strategies, see the appendix)

6. Training curriculum using mixture distributions

In more realistic scenario, the distribution of noise and other distortion parameters in the test data can be modeled as an application-dependent continuous distribution over distortion parameters. In our work, we extended the discrete case by considering a continuous distribution over Additive White Gaussian noise (AWGN) distortion. We sample noise power from the target distribution and use this parameter for sampling AWGN noise for the CIFAR-10 data set.

For evaluation of our methods, we have considered two settings for target QPDF. In the first settings, AWGN is generated with standard deviation values drawn from a continuous uniform distribution, std~*Uniform* [0,80]. For the second settings, AWGN is generated with standard deviation values drawn from a continuous gaussian distribution, std~*Normal* [40,15^2].

We choose to generate "easy" examples using Beta distribution with (a=1, scale) parameters. We assume that "simpler" examples correspond to smaller values of distortion parameter. Using the scale parameter of the Beta distribution, the probability mass around zero can be controlled by varying scale parameters (scale>1).

We construct a mixture distribution $QPDF_{\lambda,scale}(q)$ by combining the target distribution $QPDF_{\lambda=1}(q)$ with Beta distribution as described above. Essentially, we sample distortion parameter from the target distribution with probability 'λ' and from Beta distribution with probability '$1-\lambda$'. Additionally, the scale parameter of Beta distribution is used for moving the probability mass closer or away to zero.

The training curriculum is generated for each epoch by adjusting values of $'\lambda'$ and $'scale'$. Using the selected values, we sample for distortion parameters for each image in the training set.

The main challenge in curriculum learning is finding the right sequence for reaching the target $QPDF_{\lambda=1}(q)$ as training progresses. We found that linear and quadratic functions for the parameter 'λ' produce better results than the conventional method in both settings. For the parameters of Beta distribution, we found that a linear approach is well suited for our experiment.

Figure 3 shows several possible strategies for varying the parameters of mixture distribution. *Figure 4* shows few examples of the resulting distributions. In both examples (top and bottom) the distribution of the distortion parameter approaches the target distribution (uniform in this case).

In *Figure 5* we see that both linear and quadratic curriculum approaches outperform the baseline in both settings. In settings 1, both curriculum methods improve accuracy with 13% each. In settings 2, the linear approach has improved test accuracy by 12% and quadratic scheme

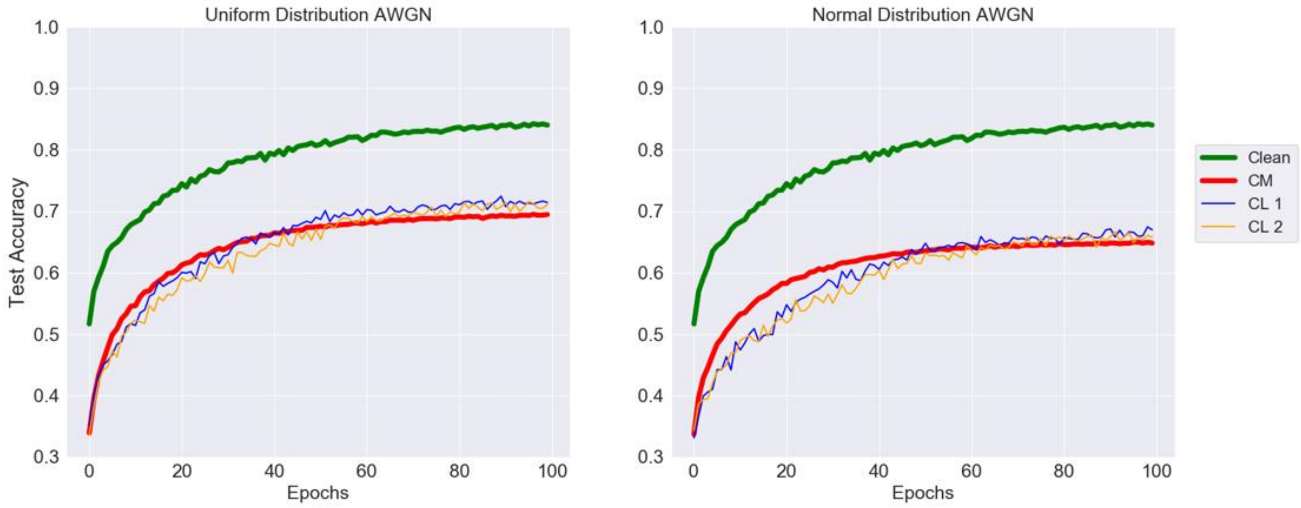

Figure 5: Test accuracy with respect to epochs. Left: test accuracy for curriculum methods and baseline in the case of uniform QPDF(q).Right: test accuracy for curriculum methods and baseline in the case of normal QPDF(q).'CL 1' refers to linear schedule while 'CL 2' refers to quadratic schedule.

by 5%. All other schedules with different $'\lambda'$ parameter (see the appendix) does not improve test accuracy over the baseline and clearly overfits.

Unlike the discrete target distribution, here all the strategies produce the same smooth behaviour, without drastic change in accuracy, as training progress. We conjecture that this is probably because the network has relatively large enough capacity to handle this task in addition to not having heavy change in the training mixture.

7. Conclusions and Future Work

In this work, we have presented a curriculum learning -based approach for training distortion-resilience neural networks. The proposed ECL method starts from an explicit probabilistic model for target distribution over distortion parameters in the test data. We have described a general recipe for building valid training curriculum using distribution mixture. The training curriculum is constructed using parameterized mixture distribution that combined the target distribution with "relaxed" distribution for generating less distorted and, thus, simpler examples. This mixture distribution allows constructing various training curricula by varying the parameters of the distribution mixture according to a different strategy.

We have demonstrated the approach with CIFAR-10 data set for image classification. The results indicate that the ECL method allows to reduce the generalization error in both discrete and continuous target distributions. However, in both cases there are issues that can arise:

There are several other ideas that can be explored for extending the ECL method. For instance, the schedule of the curriculum could be constructed adaptively. In this case, the mixture parameters should be changed adaptively, depending on the stability of the validation error archived using previous values of parameters.

References

[1] - Zhuo Chen,Weisi Lin and Shiqi Wang.2017.Image Quality Assessment Guided Deep Neural Networks Training

[2] - Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.

[3] - Fei Yang,Qian Zhang,Miaohui Wang and Guoping Qiu.2018. Quality classified image analysis with application to face detection and recognition

[4] - Samuel Dodge and Lina Karam.2018. Quality Robust Mixtures of Deep Neural Networks

[5] - Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on. IEEE, 1–6

[6] - Yu Liu, Junjie Yan, and Wanli Ouyang. 2017.Quality aware network for set to set recognition

[7] - Bengio Yoshua,Louradour Jerˆome, Collobert Ronan and Weston Jason. 2009. curriculum learning

[8] - Karen Simonyan and Andrew Zisserman. 2014.Very Deep Convolutional Networks for Large-Scale Image Recognition.arXiv preprint arXiv:1409.1556 (2014)

[9] - Fei Yang,Qian Zhang,Miaohui Wang and Guoping Qiu.2018. Quality classified image analysis with application to face detection and recognition

[10] - Xudong Sun, Pengcheng Wu, and Steven C. H Hoi.2017. Face detection using deep learning: An improved faster rcnn approach

[11] - Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof.2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark

[12] - Matthew D Zeiler. Adadelta: An adaptive learning rate method. preprint arXiv:1212.5701, 2012.

[13] - Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998

[14] - Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009

[15] - D. Arpit, S. Jastrz˛ebski, N. Ballas, D. Krueger, E. Bengio, M. Kanwal, T. Maharaj, A. Fischer,A. Courville, and Y. Bengio. A closer look at memorization in deep networks. In ICML, 2017

[16] - L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann, "Self-paced curriculum learning," in AAAI Conference on Artificial Intelligence (AAAI), 2015.

[17] - I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016

[18] - Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015b.

[19] - R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," arXiv preprint arXiv:1805.10339, May 2018.

[20] - Samil Karahan et al.,"How Image Degradations Affect Deep CNN-based Face Recognition," arXiv preprint arXiv:1608.05246, 2016.

[21] - Agostinelli F, Anderson MR, Lee H. Adaptive multi-column deep neural networks with application to robust image denoising. In: Advances in Neural Information Processing Systems;2013. p. 1493–1501

[22] - Koziarski M, Cyganek B. Deep Neural Image Denoising.In: International Conference on Computer Vision and Graphics.Springer International Publishing; 2016. p. 163–173.

[23] - Schuler CJ, Hirsch M, Harmeling S, Schölkopf B. Learning to deblur. IEEE transactions on pattern analysis and machine intelligence. 2016;38(7):1439–1451

[24] - Ghifary, M., Kleijn, W.B., Zhang, M.: Deep hybrid networks with good out-of sample object recognition. In: 2014 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). pp. 5437{5441. IEEE (2014)

[25] - Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A.,Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. CoRR,abs/1409.0575, 2014

[26] - S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS,2015.

[27] - Hossein Nejati, V Pomponiu, Thanh-Toan Do, Yiren Zhou,S Iravani, and Ngai-Man Cheung, "Smartphone and mobile image processing for assisted living," IEEE Signal Processing Magazine, pp. 30–48, 2016.

[28] - Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In:Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)

[29] - Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng,Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI,2016.

[30] - D. Weinshall, G. Cohen, and D. Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In ICML, 2018.