NOISE-ROBUST AND RESOLUTION-INVARIANT IMAGE CLASSIFICATION WITH PIXEL-LEVEL REGULARIZATION

Taesik Na, Jong Hwan Ko and Saibal Mukhopadhyay

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, GA 30332

ABSTRACT

Noise robust image classification is essential for successful deep learning deployment for the internet of things (IoTs) especially for the edge devices under stingy energy budget. Inherent image sensor noise and intentional image quality degradation for region of interest (RoI) coding reduce accuracy of image classification. In this paper, we enhance the accuracy of image classification on the perturbed images by utilizing embedding space for both image classification and additional pixel level regularization. To this end, we inject pair of clean and perturbed images during training and minimize the distance between the two resulting embeddings. We study the effects of random noise, low resolution, and mixed resolution due to RoI encoding. We experiment our algorithm for MNIST, CIFAR10 and ImageNet and show improved robustness for perturbed images compared to baseline data augmentation approach.

Index Terms— Image classification, noise, embedding, low resolution, regularization

1. INTRODUCTION

Image classification using deep learning is being widely adopted for the internet of things (IoTs) [1]. For power hungry edge devices, it is critical to manage the trade-off between energy and quality of the captured image. Region of interest (RoI) based coding is becoming a norm for controlling the energy quality trade-off in resource constraint edge devices [2]. Also, inherent image sensor noise has to be considered for successful image classification for low-end devices [3]. In this paper, we treat all these image quality degradation including low resolution (LR), random noise and mixed resolution (MR) in a single image due to RoI coding as pixel level perturbation. We aim to improve the robustness of a classifier against such perturbations.

Many prior work have been studied the impact of low quality images on the image classification. There are two ways to improve the classification accuracy. One is to remove such perturbation itself before performing classification. GSM [4] , KSVD [5] and BM3D [6] are well known algorithms for denoising and there have been several approaches using deep learning as a filter for denoising [7, 8]. Super resolution can also be applied as a pre-processing for LR images before image classification [9].

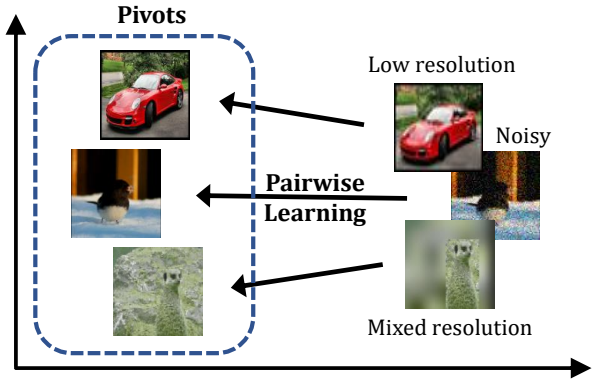

Fig. 1. Use of embedding space for learning low-level similarities between clean and perturbed images.

Another approach is to make image classifier robust against such image perturbation. [10] proposed fine-tuning a classifier with LR images after training the network with high resolution (HR) images. [11] proposed a two step training of partially coupled network for LR image classification. [12] studied the effect of noise and image quality degradation on the accuracy of a network, and proposed to re-train/fine-tune the network with data augmentation.

Our work falls in the latter category and makes the following contributions. First, we propose to utilize embedding space for both image classification and *general pixel level regularization*. Inspired by the previous work [13] where adversarial noises are regularized with a unified embedding, we use the embedding space for regularizing general pixel level perturbations including LR, random noise and mixed resolution (MR) in a single image, which were not considered in the earlier study [13].

Second, unlike [10, 11] in which they focused on accuracy for LR images only, we propose to improve accuracy of a classifier *for perturbed images while maintaining decent*

accuracy for the original images. We apply our method for MNIST, CIFAR10 and ImageNet dataset considering various pixel perturbations and show improved accuracy compared to plain data augmentation approach.

2. PROPOSED APPROACH

Simple way to improve robustness against image perturbation is to train a classifier with perturbed images in addition to original images. We advance this data augmentation approach by adding simple, yet efficient regularization [13]. When we train a classifier, we create k perturbed images using clean images and formulate a mini-batch of size m including both perturbed images and their corresponding clean images. We inject the pairs of these images during training. Un-normalized logits (embeddings) right before the softmax layer are used for both image classification and low-level pixel similarity learning. Standard cross entropy loss is used for image classification after the soft max layer and distance based loss is used for regularization of pixel value perturbation. In particular, we minimize the distance between clean and perturbed embeddings resulted from the clean and perturbed images respectively as shown in figure 1. The intuition behind the use of distance based loss is to let the classifier aware of the pairs of two embeddings are from the same original images. This additional regularization promotes the classifier to ignore the pixel noises (difference between perturbed images and the original images). The entire procedure is described in algorithm 1.

We formulate the total loss as follows.

$$Loss = \frac{1}{(m-k)+\lambda k}(L_{org}+\lambda L_{perturb})+\lambda_2 L_{dist} \quad (1)$$

where, L_{org} is the classification loss for the original images, $L_{perturb}$ is the classification loss for the perturbed images, L_{dist} is the distance based loss between the pairs of original and perturbed images. m is the size of the mini batch, k ($\leq m/2$) is the number of perturbed images in the mini batch. λ is the parameter to control the relative weight of the classification loss for perturbed images. λ_2 is the parameter to control the relative weight of the distance based loss L_{dist} in the total loss. Each loss term is defined as follows:

$$L_{org} = \sum_{i=1}^{m-k} L(\boldsymbol{X}_i|y_i)$$

$$L_{perturb} = \sum_{i=1}^{k} L(\boldsymbol{X}'_i|y_i)$$

$$L_{dist} = \sum_{i=1}^{k} ||\boldsymbol{E}'_i - \boldsymbol{E}_i||_2^2$$

where, $\boldsymbol{X}'_i$ is i'th perturbed image generated from i'th original image $\boldsymbol{X}_i$. $\boldsymbol{E}_i$ and $\boldsymbol{E}'_i$ are the resulting embeddings from $\boldsymbol{X}_i$ and $\boldsymbol{X}'_i$ respectively. We use pivot loss introduced in [13] where the gradient back propagation is not performed through original embeddings $\boldsymbol{E}$. Additional details can be found in [13].

Algorithm 1 Image classification with pixel level regularization

m: size of mini batch, k: size of perturbed images

Require: $k \leq m/2$

1: **repeat**
2: Get mini batch $B = \{X_1, ..., X_{m-k}\}$.
3: Generate k perturbed examples $\{X'_1, ..., X'_k\}$ from corresponding original examples $\{X_1, ..., X_k\}$.
4: Make new mini batch $B' = \{X'_1, ..., X'_k, X_1, ..., X_{m-k}\}$.
5: Perform one step forward pass with B'.
6: Formulate the cross entropy loss with entire embeddings $\{E'_1, ..., E'_k, E_1, ..., E_{m-k}\}$.
7: Formulate the distance based loss with perturbed embeddings $\{E'_1, ..., E'_k\}$ and corresponding original embeddings $\{E_1, ..., E_k\}$.
8: Perform one step backward pass and update the parameters in N.
9: **until** training converged

3. EXPERIMENTAL RESULTS

3.1. MNIST and CIFAR10

We use 20-layer ResNet (table 6 in [14]) model for MNIST and CIFAR10 dataset. We scale down the image values to [0,1] for both dataset and subtract per-pixel mean values for CIFAR10. We perform 32x32 random crop and random flip on zero padded 40x40 original images for CIFAR10. We train networks with λ = 0.3, λ_2 = 0.0001, m = 128, k = 64 for the experiments. Stochastic gradient descent (SGD) optimizer with momentum of 0.9, weight decay of 0.0001 are used. We start training with a learning rate of 0.1 and divide it by 10 at 4k, 6k and 8k iterations and terminate training at 9k iterations for MNIST, and 48k and 72k iterations, and terminate training at 94k iterations for CIFAR10.

3.1.1. Random Noise

We use Gaussian noise as an example of random noise. During training, we generate Gaussian noise $\mathcal{N}(\mu = 0, \sigma^2)$ added images where σ is selected randomly in the interval [0, max_σ] per each image. We clip the noise added images with the range [0,1]. We train networks injecting those noisy images and clean images with/without pivot loss. We also include standard training with clean images and only with noisy images for comparison.

Table 1 shows test accuracy for MNIST and CIFAR10

Table 1. Test accuracy (%) for Gaussian noise on MNIST and CIFAR10 dataset. (Clean, Noisy / Pivot loss) are trained with max_σ = 0.15. Higher accuracy among (Ours) and (Clean, Noisy) is emphasized in **bold**.

MNIST

Training	clean	σ = 0.15
Clean only	99.6	40.9
Noisy only	99.5	99.5
Clean, Noisy	**99.6**	99.4
Pivot loss (Ours)	99.5	**99.5**

CIFAR10

Training	clean	σ = 0.05	σ = 0.15
Clean only	91.6	62.4	16.6
Noisy only	89.8	88.9	82.0
Clean, Noisy	90.5	89.1	80.7
Pivot loss (Ours)	**90.8**	**89.6**	**81.7**

dataset. As expected, we observe injecting noisy images during training improves accuracy for noisy images at test time compared to training only with clean images (Clean only) for both MNIST and CIFAR10. For simple images like MNIST, our pivot loss doesn't show any meaningful difference compared to data augmentation approach (Clean, Noisy) or training with noisy images only case (Noisy only).

Training only with noisy images: for CIFAR10 dataset, training only with noisy images (Noisy only) shows best accuracy for noisy images with higher σ (σ=0.15), however, at the expense of decreased accuracy for the clean images. It is natural that the network trained with noisy images perform well for the noisy images, but not for the clean images. Careful selection of σ is necessary when we use Gaussian noise as a data augmentation for the clean images since larger noise actually decreases accuracy for the clean images.

Training with clean and noisy images: (Clean, Noisy / Pivot loss) show good compromise between training only with clean images (Clean only) and with noisy images (Noisy only). Our pivot loss only increase accuracy for both clean and noisy images compared to pure data augmentation approach. This shows that additional loss with a unified embedding results in better regularization than training only with augmented data when we deal with pixel level perturbation.

3.1.2. Low Resolution

We augment LR images with sub-sampling factors of either 2 or 4. After down-sampling, those images are up-sampled with 'nearest' or 'bicubic' method. We randomly choose sub-sampling factor and up-sampling method per each image during training. Resulting LR images together with the clean high resolution (HR) images are injected during training with/without pivot loss. Again, we also train networks with clean images and with LR images for comparison.

Table 2. Test accuracy (%) for LR images (sub-sampling = x4) on MNIST and CIFAR10 dataset. For CIFAR10 dataset, 110-layer ResNet models are also trained to analyze the effect of pivot loss for deeper networks. Higher accuracy among (Ours) and (Clean, LR) is emphasized in **bold**.

MNIST

Training	clean	x4 (nearest)	x4 (bicubic)
Clean only	99.6	50.4	57.5
LR only	99.5	87.6	86.4
Clean, LR	**99.6**	86.8	86.0
Pivot loss (Ours)	**99.6**	**87.1**	**86.2**

CIFAR10

Training	clean	x4 (nearest)	x4 (bicubic)
20-L, Clean only	91.6	19.2	19.9
20-L, LR only	86.9	75.9	72.9
20-L, Clean, LR	91.7	71.5	70.7
20-L, Pivot loss (Ours)	**92.0**	**73.5**	**72.7**
110-L, Clean only	93.5	21.9	15.7
110-L, LR only	86.4	78.8	77.0
110-L, Clean, LR	**93.8**	76.4	76.2
110-L, Pivot loss (Ours)	93.3	**78.1**	**77.4**

Table 2 shows test accuracy of the trained networks for clean and LR images. In this case as well, as expected, we observe training with LR images improves accuracy for LR images compared to training only with clean images (Clean only) on both MNIST and CIFAR10.

Training only with LR images: for MNIST, training only with LR images (LR only) results in good accuracy for clean images as well as LR images. This is because MNIST LR images have good enough features that can be well generalized for clean HR images. For complex images like CIFAR10, however, we observe training only with LR images (LR only) degrades accuracy for clean HR images since LR images lose features that are necessary for HR image classification. In all cases, accuracy for LR images is not close to that for clean images due to the loss of features in LR images.

Training with clean and LR images: interestingly, networks trained with both clean and LR images (Clean, LR / Pivot loss) show better accuracy improvements for both clean and LR images compared to the network trained with clean images (Clean only). This suggests that data augmentation with LR images serves as a good regularizer for the clean images as well. However, (Clean, LR / Pivot loss) show decreased accuracy for LR images compared to (LR only). This

Fig. 2. Generation of mixed resolution (MR) in a single image. Blue box and red box in the original image represent random crop area and bounding box respectively.

is because the network trained only with LR images has seen more LR images during training, thus, performs better for LR images. Again, our pivot loss only increases accuracy for clean and LR images compared to the network trained without pivot loss.

Deeper networks: we further test the accuracy for 110-layer ResNet models on CIFAR10 dataset to study the effect of training both with clean and LR images for deeper networks. As seen from the table 2, data augmentation with 110-layer ResNet improves accuracy for both original and LR images compared to 20-layer counter part. Our pivot loss increases accuracy for LR images compared to data augmentation approach, however, not for the clean images. The experiments show that our pivot loss is always a good regularizer for LR images, however, it is sometimes good or bad regularizer for the clean images. Thus, we recommend using distance based loss when the focus is to improve the accuracy of LR images while maintaining good enough accuracy for the original images. We also recommend using distance based loss with LR images for fine-tuning on the accuracy for original images as observed in 20-layer ResNet case study.

3.2. ImageNet

We use 18-layer ResNet (table 1 in [14]) model for ImageNet [15] classification. We scale down the image values to [0,1] and subtract per-channel mean values. We perform random crop ($0.5 <$ area range < 0.8) and random flip on original images, and resize images to 244x244. We start training with a learning rate of 0.1 and divide it by 10 at 500k, 800k and 1M iterations and terminate training at 1.1M iterations. We first train a network only with clean images.

3.2.1. Mixed Resolution in a Single Image

To see the effect of region of interest (RoI) based encoding on image classification, we consider mixed resolution (MR) in a single image as in figure 2. We use ground truth bounding boxes to create MR images for training, and use region-based fully convolutional network (R-FCN) [16] with ResNet-101 trained on MS-COCO dataset [17] for bounding box generation on validation data set at test time. [1] We fine-tune the baseline network with clean, MR and LR images for 400k iterations with a learning rate of 1e-5 with/without pivot loss. We use $\lambda = 0.3$, $\lambda_2 = 0.0001$, $m = 64$ and $k = 32$ (16 for LR, 16 for MR) for the experiments.

Table 3. Test accuracy (%) for MR and LR images (subsampling = x4) on ImageNet dataset. Higher accuracy among (Ours) and (Clean, MR, LR) is emphasized in **bold**.

Training	clean	MR x4 (nearest)	LR x4 (nearest)
Clean only	68.1	57.8	29.3
Clean, MR, LR	**67.9**	61.5	44.9
Pivot loss (Ours)	67.4	**62.1**	**49.0**

Table 3 shows accuracy results for MR and LR images. Pivot loss shows better accuracy for MR and LR images compared to data augmentation approach (Clean, MR, LR) at the expense of small accuracy decrease on the clean images. This shows that pivot loss serves as a good regularizer for images with various resolution including MR and LR on ImageNet dataset. This result is also consistent with the LR case study for deeper networks on CIFAR10 dataset. As long as resolution and noise are concerned for image classification, our pivot loss can be a simple, yet efficient regularizer for perturbed images while maintaining decent accuracy for the clean images.

4. CONCLUSION

We proposed the use of embedding space for both image classification and pixel level regularization for various types of image perturbation. We considered random noise, LR and MR due to RoI based coding that can be happen in practical IoT scenario as examples of image perturbation. Image classification results on MNIST, CIFAR10 and ImageNet dataset showed promising results when we used proposed pivot loss as an additional regularization compared to the baseline data augmentation approach.

Acknowledgments

The research reported here was supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-17-2-0045. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA.

[1] Since ImageNet validation dataset doesn't have bounding boxes, we use R-FCN just for bounding box generation and observe decent quality of bounding box generation.

5. REFERENCES

[1] Jong Hwan Ko, Burhan Ahmad Mudassar, Taesik Na, and Saibal Mukhopadhyay, "Design of an energy-efficient accelerator for training of convolutional neural networks using frequency-domain computation," in *Design Automation Conference, (DAC)*, 2017, pp. 59:1–59:6.

[2] Jong Hwan Ko, Taesik Na, and Saibal Mukhopadhyay, "An energy-efficient wireless video sensor node with a region-of-interest based multi-parameter rate controller for moving object surveillance," in *Advanced Video and Signal Based Surveillance, (AVSS)*, 2016, pp. 138–144.

[3] Jong Hwan Ko, Mohammad Faisal Amir, Khondker Zakir Ahmed, Taesik Na, and Saibal Mukhopadhyay, "A single-chip image sensor node with energy harvesting from a CMOS pixel array," *IEEE Trans. on Circuits and Systems I*, vol. 64-I, no. 9, pp. 2295–2307, 2017.

[4] Javier Portilla, Vasily Strela, Martin J. Wainwright, and Eero P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. Image Processing*, vol. 12, no. 11, pp. 1338–1351, 2003.

[5] Michael Elad and Michal Aharon, "Image denoising via learned dictionaries and sparse representation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 895–900.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[7] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 350–358.

[8] Harold Christopher Burger, Christian J. Schuler, and Stefan Harmeling, "Image denoising: Can plain neural networks compete with bm3d?," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2392–2399.

[9] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] Xingchao Peng, Judy Hoffman, Stella X. Yu, and Kate Saenko, "Fine-to-coarse knowledge transfer for low-res image classification," in *2016 IEEE International Conference on Image Processing, (ICIP)*, 2016, pp. 3683–3687.

[11] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S. Huang, "Studying very low resolution recognition using deep networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Yiren Zhou, Sibo Song, and Ngai-Man Cheung, "On classification of distorted images with deep convolutional neural networks," in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2017, pp. 1213–1217.

[13] Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay, "Cascade adversarial machine learning regularized with a unified embedding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li, "Imagenet: A large-scale hierarchical image database," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[16] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-FCN: object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 379–387.

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, vol. 8693, pp. 740–755.