Towards End-to-End Low-resolution Image Classification without Super-Resolution Network by Meta-Learning on Downsampling and Quantization

Jiseok Youn, Youngseok Kim, Jayu Kim, Myeonggeun Jun
Seoul National University
{jsjs0369, kyssnu, ,rlawkdb1225, wjsaud0909}@snu.ac.kr

Abstract

This paper presents a new possibility of more efficient low-resolution image classification via end-to-end training without a super-resolution network. We consider a low-resolution image as downsampled and quantized version of corresponding high-resolution image. At meta-training, the proposed model learns to learn how to classify low-resolution images which has been converted from high-resolution by a possible combination of downsampling and quantization. After meta-training, the model quickly adapts to classify low-resolution images degraded in an unseen way, by fine-tuning with a few labeled low-resolution examples. Our scheme is more lightweight compared to existing works that need to train and infer a neural network just for super-resolution. The experiment results with Food-101 on CNN show that our method can increase classification accuracy without a demand of super-resolution network.

1. Introduction

The image classification accuracy of Convolutional Neural Network (CNN) is greatly affected by the resolution of the input image. There are cases when it is necessary to use low-resolution images, such as small original size medical images or images taken from drones at high altitudes. Therefore, low-resolution image classification should be studied to implement the real-world image classification.

Lots of methods have been proposed to classify the low-resolution images. Most of the research implemented the Super-Resolution (SR) method which converts the low-resolution images into the high-resolution images. Wang et al. proposed an attribute embedded discriminative network to super-resolve very low-resolution images [25]. In order to re-identify a person, a high-resolution probe image is classified from a gallery set which composed of the low-resolution images. By adjusting the scale through the generator network, the limitation that the gallery set images do not have uniform size has been solved. Jiao et al. integrated the SR sub-network and re-identification sub-network to improve the integration compatibility [9]. This method has the advantage of reducing the computational load of SR and improving the performance through end-to-end joint optimization. However, it did not solve the fundamental issue of the SR network. Zhou et al. introduced a weight map representing the positions of pixels containing high-frequency information in the real high-resolution image [29]. A pixel-level loss function is used to reduce the errors between the ground-truth high-resolution images and predicted images.

Since the above studies are accompanied by the network-like SR module, it is difficult to apply them to devices in low-computation such as smartphone. Li et al. designed a semi-coupled projective dictionary learning to re-identify the low-resolution image without SR [13]. Singh et al. applied a capsule network which considers the properties of objects to the Very Low-Resolution (VLR) images [26]. Since performance deteriorates when VLR is implemented only using a CapsNet, low-resolution images were classified through the unlabeled high-resolution images.

In this paper, we propose a new method for efficient low-resolution image classification excluding a heavy SR network and utilizing meta-learning. Meta-learning enables quick adaptation into an unseen task by leveraging past experience on different tasks.

The main contributions of the proposed algorithm are as follows:

- The proposed scheme uses meta-learning where a task is defined as a way of image degradation into low-resolution. During meta-training, the proposed scheme converts input images via some possible combinations of downsampling (e.g., max pooling, average pooling) and quantization (e.g., quantization into 8-bit).

- The proposed scheme is favorable for mobile deployment since the SR sub-network is not involved to the model. There is less additional overhead on storage and computation than the complicated SR-based approach requires.

2. Related works

2.1 Meta-learning

Meta-learning can be expressed as 'learning to learn'. It enables quick adaptation into an unseen task by leveraging past experience on diverse (somewhat) related tasks. Previous meta-learning works can be split into optimization-based [5], metric-based [20], and model-based [23]. Among them, we use Model-Agnostic Meta-Learning (MAML), a pioneering and representative optimization-based meta-learning. Most meta-learning works are for few-shot learning application which aims to quickly adapt to images with unseen classes and to classify them correctly. Therefore, a task is usually defined as a set of interested classes. However, as said earlier, we introduce a new definition of task since our scheme is for low-resolution image classification without a SR network.

2.2 Quantization

Quantization means converting analog data into digital data. General quantization reduces the number of bits to be used when expressing digital data, thereby reducing the model size. Many quantization methods are being studied to apply quantization to deep learning model weight compression. There are a method of quantizing the weight of an already trained model [7] and a method of matching the weight value to the value to be quantized while training the model [1, 4]. Among the methods of quantizing the weights of an already trained model, research (mixed precision) [17, 22, 24] to find different optimal bits for each layer is being actively conducted in the image classification field. This method minimizes performance degradation by quantizing the pre-learned weights into different bits for each layer. There is a method that solves the problem in a differentiable way [8], and there are HAQ [24] and AutoQ [17] that use reinforcement learning. HAQ [24] found the optimal bit for weights and activations using DDPG [14], and AutoQ [17] found the optimal bit for kernels and activations using HIRO [18], which hierarchically uses reinforcement learning agents.

2.3 Super-resolution

SRCNN [2] is the first model that performed Super-Resolution (SR) using CNN. Though it is a relatively simple model made by stacking only three layers, It surpasses the existing traditional machine learning-based SR performance. SRCNN has proposed a method of increasing the size of a low-resolution image by linear interpolation and passing the enlarged image through the CNN to obtain a restored image. Because it passed the enlarged image through the CNN, there was a disadvantage of consuming a lot of computing power. It had limitations in terms of accuracy as it was a simple structure using three CNN layers. In methods such as FSRCNN [3] and ESPCN [19] proposed later, unlike SRCNN, the LR image is put in the CNN input and then the size is enlarged in the output layer to reduce the computing power and increase the CNN layer. However, as the layers of the CNN deepen, a problem of vanishing gradient occurred, in which the information in the front layer was gradually lost as it passed through the layers during training. VDSR [11] improved the performance than SRCNN by using 20 layers while introducing a residual learning technique using skip connection. In addition, deeper models [12, 16, 28] with better performance were proposed by applying this, but they did not take into account the model inference time, such as using 800 layers [28]. In addition, as new CNN techniques, DRN [6], USRNet [27], MZSR [21], etc. have been proposed. Unlike the conventional model that generally uses one loss value, DRN adds a dual regression loss to the existing loss and combines the two loss values and uses it as a loss function. The dual regression loss is limited to be similar to the input LR image when downsampling is performed on the reconstructed image from the LR. USRNet and MZSR methods are models proposed to perform robust super-resolution in various kernel environments. USRNet is a method to restore an image by setting the noise level and kernel type as hyperparameters and adjusting them.

3. Method

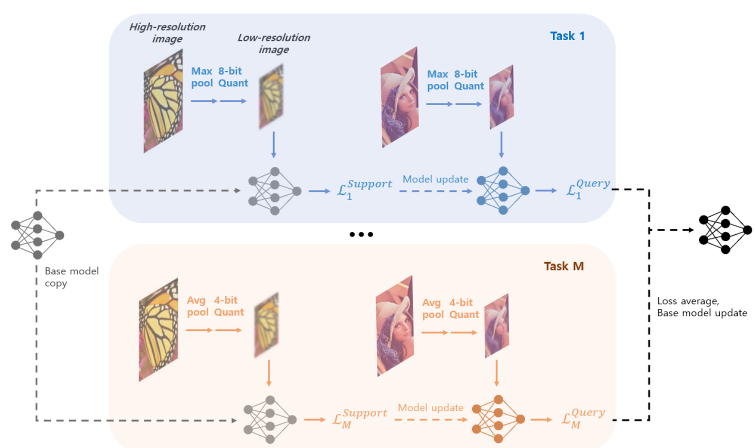

Figure 1. Overview of proposed meta-learning scheme.

Figure 1. shows the overview of our proposed scheme. We divide the process of generating a low resolution image from a high resolution image into two operations: downsampling and activation quantization. Our interest is on a base model that needs to be adaptive to any task. Optimiza-

Layer Name	Output Size (w/ SR or w/o SR)	
Training input (raw)	Bx3x32×32	Bx3x128×128
Training input (resized)	Bx3x128×128	Bx3x32×32
Conv3x3-BN-ReLU 1	Bx32x128×128	Bx32x32×32
Max pooling 1	Bx32x64×64	Bx32x16×16
Conv3x3-BN-ReLU 2	Bx32x64×64	Bx32x16×16
Max pooling 2	Bx32x32×32	Bx32x8×8
Conv3x3-BN-ReLU 3	Bx32x32×32	Bx32x8×8
Max pooling 3	Bx32x16×16	Bx32x4×4
Conv3x3-BN-ReLU 4	Bx32x16×16	Bx32x4×4
Max pooling 4	Bx32x8×8	Bx32x2×2
Flatten	Bx2048	Bx128
fc, softmax	Bx101	

Table 1. Structure of compared (left, with SR) and proposed (right, without SR) CNNs. Resize operation is done by SR network or selected downsampling-quantization. B stands for mini-batch size.

tion of the base model requires M query (post-adaptation) loss from M meta-tasks. In every meta-task, our scheme randomly selects downsampling method and quantization method among two predefined set, respectively. A meta-task needs two labeled mini-batches; support (to adapt via repetitive fine-tuning) and query (to evaluate the adaptation), which are converted from high-resolution by the combination of selected downsampling and quantization methods (i.e., the task). The base model is adapted to task-specific model by fine-tuning with support mini-batch, then the task-specific model outputs query loss with query mini-batch. From the averaged query loss over M meta-tasks, the base model is updated (one 'step').

4. Experiments

4.1. Compared scheme

We adopt Enhanced Deep Super-Resolution (EDSR) [15] as the compared scheme using SR network. EDSR has enhanced SR performance by removing unnecessary modules in conventional residual networks and expanding the model size along with training stabilization. We've fetched EDSR using the official PyTorch code uploaded on Github. EDSR is before the CNN model for image classification. We use pretrained EDSR (using DIV2K) then freeze it. In other words, training with Food-101 dataset is only for updating the CNN for classification (see Table 1).

4.2. Proposed scheme

We've implemented the proposed meta-learning scheme using PyTorch. SR network doesn't exist, but there are $M = 4$ meta-tasks per step. In every meta-task, downsampling method is selected among {max pooling, average pooling}. Meanwhile, in every meta-task, quantization method on every pixel value is also selected among {2, 3, 4, 8, 16, 32}-bit. For quantization, we use the operation on activation in DoReFa-Net [30], an early work on quantization-aware training. The number of repetitive fine-tuning with a support mini-batch equals to 5. Because there is no SR network, training with Food-101 dataset is only for updating the base CNN for classification (see Table 1).

4.3. Dataset

We mainly use Food-101 dataset [10] for meta-learning, fine-tuning, and inference. It consists of 101 food categories with 750 training and 250 test images per category, making a total of 101k images. The labels for the test images have been manually cleaned, while the training set contains some noise.
We do not directly use DIV2K dataset, but it has been used for pretraining EDSR. It consists of 1000 2K resolution RGB images which contain a large diversity of contents. These images are divided into three parts: 800 images for training, 100 images for validation, and 100 images for testing.

4.4. model architecture

In similar to [5], we use 4 convolution blocks including 3x3 convolution using 32 filters, Batch Normalization (BN), and ReLU activation function. Table 1 shows the structure of compared (left, with SR) and proposed (right, without SR) CNNs.

4.5. Training details

For the compared scheme using EDSR and CNN, we use Adam optimizer with learning rate 0.01. On the other hand, for the proposed scheme using CNN, we use Adam optimizer with learning rate 0.0001 for base model optimization and SGD optimizer with learning rate 0.01 for task-specific model optimization. Batch size is set to 32 for training and 16 for validation. In validation, regardless of compared/proposed scheme, an input is of size Bx3x32x32, degraded by bicubic operation (neither SR nor downsampling-quantization). The total number of epochs equals to 11.

4.6. Result

We can see that it is possible to reduce the training and validation loss even without a heavy SR network. However, due to the insufficient hyper-parameter tuning and (meta-)training time, we couldn't compare the converged performance.

5. Future work and Conclusion

We've explored a new possibility of more efficient low-resolution image classification via end-to-end training without a super-resolution network. We regard a low-resolution

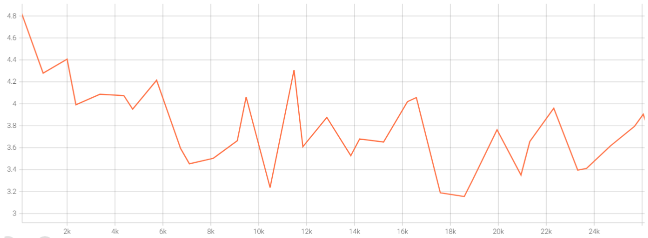

Figure 2. Training loss vs. steps for EDSR-attatched CNN. There are 2368 steps per epoch.

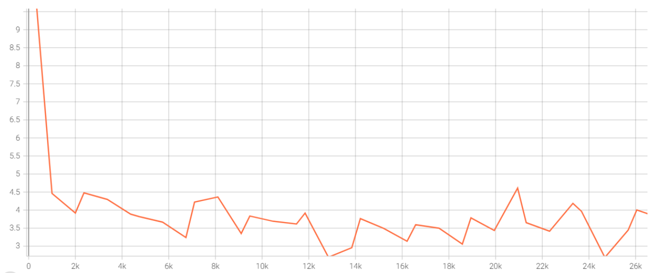

Figure 3. Validation loss vs. steps for EDSR-attatched CNN. There are 2368 steps per epoch.

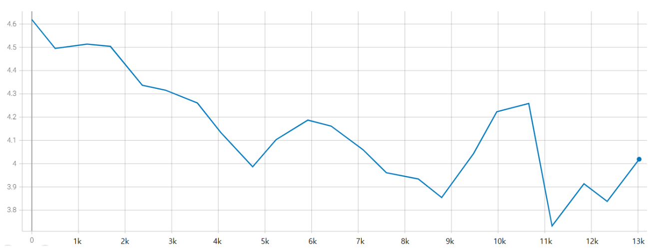

Figure 4. Training query loss averaged over $M = 4$ tasks vs. steps for proposed CNN. There are 1184 steps per epoch.

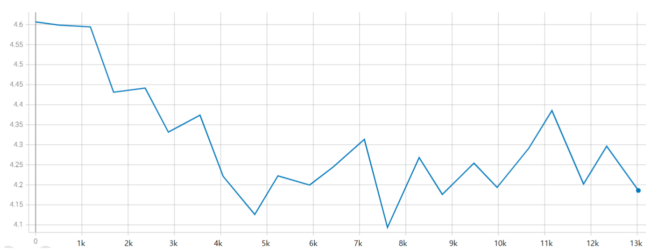

Figure 5. Validation query loss vs. steps for proposed CNN. There are 1184 steps per epoch.

image as downsampled and quantized version of corresponding high-resolution image. Assuming that single CNN in a device is to classify low-resolution images degraded by a certain combination of downsampling and quantization methods, we utilize meta-learning with newly-defined task for a base model adaptive to any low-resolution or a combination of the methods. By analyzing converged performance and dealing with some remaining problems, it may be possible to develop a low-resolution CNN which really works well without SR network.

References

[1] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 2

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[3] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 2

[4] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019. 2

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. 2, 3

[6] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5407–5416, 2020. 2

[7] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 2

[8] Zejiang Hou and Sun-Yuan Kung. Efficient image super resolution via channel discriminative deep neural network pruning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3647–3651. IEEE, 2020. 2

[9] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[10] Parneet Kaur, Karan Sikka, and Ajay Divakaran. Combining weakly and webly supervised learning for classifying food images. *CoRR*, abs/1712.08730, 2017. 3

[11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[12] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 517–532, 2018. 2

[13] Kai Li, Zhengming Ding, Sheng Li, and Yun Fu. Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1

[14] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 2

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 3

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2

[17] Qian Lou, Feng Guo, Lantao Liu, Minje Kim, and Lei Jiang. Autoq: Automated kernel-wise neural network quantization. *arXiv preprint arXiv:1902.05690*, 2019. 2

[18] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:1805.08296*, 2018. 2

[19] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[20] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087, 2017. 2

[21] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020. 2

[22] Stefan Uhlich, Lukas Mauch, Fabien Cardinaux, Kazuki Yoshiyama, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Mixed precision dnns: All you need is a good parametrization. *arXiv preprint arXiv:1905.11452*, 2019. 2

[23] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638, 2016. 2

[24] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8612–8620, 2019. 2

[25] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, volume 1, page 4, 2018. 1

[26] Yuanwei Wu, Ziming Zhang, and Guanghui Wang. Unsupervised deep feature transfer for low resolution image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[27] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3217–3226, 2020. 2

[28] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2

[29] Liguo Zhou, Guang Chen, Mingyue Feng, and Alois Knoll. Improving low-resolution image classification by super-resolution with enhancing high-frequency content. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1972–1978. IEEE, 2021. 1

[30] Shuchang Zhou, Zekun Ni, Xinyu Zhou, He Wen, Yuxin Wu, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *CoRR*, abs/1606.06160, 2016. 3