Optical Character Recognition Guided Image Super Resolution

Philipp Hildebrandt*
Hasso-Plattner-Institut
University of Potsdam, Germany
philipp.hildebrandt@student.hpi.de

Maximilian Schulze*
Hasso-Plattner-Institut
University of Potsdam, Germany
maximilian.schulze@student.hpi.de

Sarel Cohen
The Academice College of Tel Aviv-Yaffo
Tel Aviv-Yafo, Israel
sarelco@mta.ac.il

Vanja Doskoč
Hasso-Plattner-Institut
University of Potsdam, Germany
vanja.doskoc@hpi.de

Raid Saabni
The Academic College of Tel Aviv-Yaffo, Triangle R&D Center
Tel-Aviv, Israel
raidsa@mta.ac.il

Tobias Friedrich
Hasso-Plattner-Institut
University of Potsdam, Germany
tobias.friedrich@hpi.de

ABSTRACT

Recognizing disturbed text in real-life images is a difficult problem, as information that is missing due to low resolution or out-of-focus text has to be recreated. Combining text super-resolution and optical character recognition deep learning models can be a valuable tool to enlarge and enhance text images for better readability, as well as recognize text automatically afterwards. We achieve improved peak signal-to-noise ratio and text recognition accuracy scores over a state-of-the-art text super-resolution model TBSRN on the real-world low-resolution dataset TextZoom while having a smaller theoretical model size due to the usage of quantization techniques. In addition, we show how different training strategies influence the performance of the resulting model.

KEYWORDS

optical character recognition, image super-resolution, deep learning, unfocused images

ACM Reference Format:
Philipp Hildebrandt, Maximilian Schulze, Sarel Cohen, Vanja Doskoč, Raid Saabni, and Tobias Friedrich. 2022. Optical Character Recognition Guided Image Super Resolution. In *Proceedings of The 22th ACM Symposium on Document Engineering (DocEng)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

When taking photos of text, information might get lost due to an insufficient resolution, wrong focus, or shaking of the camera.

Image super-resolution is concerned with increasing the resolution of low-resolution images without losing information or increasing noise. State-of-the-art image super-resolution models decode visual information and up-sample the extracted feature vectors.

In addition to increasing the picture resolution, it is also beneficial to detect text so that it can be read out without the need to manually look for text in fine details. Optical character recognition in natural scenes showed vast progress in the last time. Transformers are using a self attention mechanism [15] on semantic, positional and visual information [18] to extract text from natural scenes.

Combining both approaches might increase the results of optical character recognition on low-resolution images and distant text. As current deep learning models include many layers and computation, it is essential to reduce the model size and computation cost to a minimum while keeping the performance.

To reduce the model size and computational cost, quantization exploits the fact that 32-bit floats catch more information than needed and thus reduces the bit-width of parameters.

This work combines a state-of-the-art super-resolution network specifically trained for scene text images and a state-of-the-art scene text recognition model to maximize accuracy on the TextZoom dataset. Also, post-training quantization is used and analyzed to reduce the model size as much as possible without degrading the performance.

We provide the code for this work in a Google Drive folder. Please follow the provided link[1].

2 RELATED WORK

Image super-resolution (SR) made major progress in the last few years. The first method of SR used interpolation to increase the resolution of images. However, interpolation introduces a lot of noise into images and does not include semantic information from the image. With the introduction of deep learning architectures like convolutional neural networks (CNN) or Laplacian feature pyramids into super-resolution, the precision with which images were resized increased [6, 14]. One of the latest improvements combined an enhanced residual neural network (EDSR) [12] with a Resampler Network into a Content-Aware Resampler (CAR) [13]. This way, low-resolution representations of high-resolution images are learned, which in turn help to further train the SR model. Their SR model achieves top performance in two times upscaling on

*Both authors contributed equally to this research.

[1] https://drive.google.com/drive/folders/1L0Q1W1Rr4lIcuMpzNXVV5452FMJfj4Om?usp=sharing

DocEng, September 20th–23rd, 2022, San Jose, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

multiple common image super-resolution datasets like Set5 [3], Set14 [17] or Urban100 [10]. A subfield of image super-resolution is text super-resolution (TSR). By focusing on reconstructing high-resolution text images, the model training and design can be adapted to the characteristics of text images. Many approaches for TSR aim at capturing sequential text information by implementing recurrent neural networks or transformer layers in their TSR model [4, 16] for enhanced text clarity. Others specifically design the loss to focus on text reconstruction. One version of this is the so-called text focus loss [4]. Besides using the mean squared error of the high-resolution image and the low-resolution image, a transformer-based OCR model is trained, based on which two new terms to the loss function are added:

- The position loss is computed by utilizing the attention maps from the transformer. The attention maps are used to compute the L1 loss between the pixels containing the letters in the high-resolution image and the SR image.
- The recognition loss is based on a modified cross-entropy loss from the text prediction of the transformer OCR model and the text annotation of each image.

Using this loss both improved the SR model trained with this loss aswell as the optical character recognition (OCR) accuracy on the SR images.

OCR is one of the many text-related computer vision challenges which are tackled using deep learning. Segmentation-free scene text recognition tries to capture text in images as a whole by text to an image. Nowadays, this area can be divided into connectionist temporal classification-based methods, and attention-based methods [5]. Many new proposed scene-text-recognition methods are attention-based to capture positional information without the limitations of a short-term memory [9, 18].

One way to reduce model size is to quantize parameters. To quantize parameters after training, different methods have been proposed. Uniform quantization evenly divides the range of possible values into a given range, which is often set to the minimum and maximum values of the weights, in buckets, to which each value is assigned [7]. Another method, Analytical Clipping for Integer Quantization, defines an optimal clipping range per bit-width, arguing that weight values are often not uniformly distributed, but have a gaussian or laplacian distribution [2].

3 DATASET

We use the TextZoom dataset [16]. This dataset contains 21740 low-resolution and high-resolution image pairs of text, which all are annotated with the text depicted in each image. The high-resolution images have a resolution of 32 x 128, and the low-resolution images have a resolution of 16 x 64. They were created by capturing the same subject as in the high-resolution image with a different focal length. Depending on how much the focus was changed compared to the high-resolution image, the authors of TextZoom divided the images into easy, medium, and hard subsets with 1619, 1411, and 1343 test samples, respectively. The authors of the scene-text-telescope [4] split the dataset into 17367 train low and high-resolution image pairs and test low and high resolution 4373 image pairs. Two examples of low and high-resolution image pairs from the hard test set can be found in Figure fig. 2.

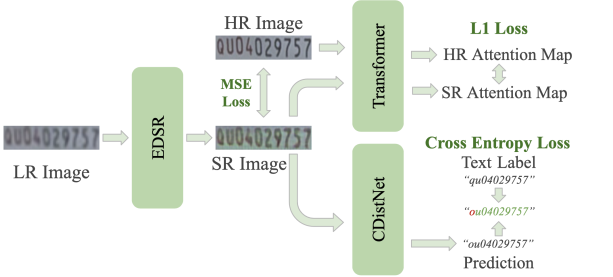


Figure 1: Modified architecture as proposed by Chen et al. [4] with substituted models.

4 METHODOLOGY

In this section, we first explain our model architecture, which consists of an image super-resolution model and an optical character recognition model, and how we train both models. Afterwards, we illustrate how we quantize both models.

4.1 Architecture

To perform OCR on the out-of-focus low-resolution images, we combine an OCR model with an Image Super-Resolution (SR), as proposed by [4]. The architecture is depicted in Figure 1. The goal of this step is to improve the OCR performance with this additional model compared to simple interpolation algorithms.

4.1.1 Image Super-Resolution Model. Instead of using an SR model specifically designed for text images, we follow a different approach by using an SR model, which was trained the DIV2K image super-resolution dataset [1] and evaluated on multiple commonly used SR datasets like Set5 [3], Set14 [17], or Urban100 [10]. We reuse the knowledge gained from non-text images, which are captured in the trained model, and fine-tune it on our dataset to re-purpose this knowledge for improved text super-resolution. Instead of changing the model architecture to optimize for text images, we rely on the text-focused loss [4], explained in the previous section, to optimize the SR images for OCR.

As SR model, we choose an EDSR for image super-resolution [12], which was trained together with a Resampler Network, that produces learned intermediate low-resolution images as proposed by [13], where they were able to train on high-resolution images and improve the EDSR. Our configuration of this network uses 32 Res-Blocks with 256 channels to double the image resolution.

We will use the previous best model on the TextZoom dataset TBSRN [4] as a comparison throughout this paper, which first feeds the low-resolution images into a Spatial Transformer Network (STN) [11] for text orientation correction. For a fair comparison, we added the same STN as the first step to our model and used the output of the STN as input for the EDSR.

4.1.2 Optical Character Recognition Model. For the optical text recognition, we use the CDistNet model [18]. The model uses a classic encoder-decoder architecture. The encoder uses three computational branches to process semantic, visual, and positional information. The visual branch uses a ResNet50 [8] and a three-layer transformer unit with self-attention to build a visual feature map. The semantic branch captures information about previously seen characters of the sequence. The positional branch embeds the position of a character in the sequence into the model.

For the decoder, they proposed a block named *MDCDP*. It uses multi-head self-attention [15] to enhance the positional features. These are then used in a cross-branch interaction to query the semantic and visual branches, which are then combined to get a position-aware semantic-visual embedding. These *MDCDP* blocks are stacked multiple times to enhance the gained information [18].

4.1.3 Training. Our training process consists of two separate steps: First, we train our SR model on the training subset of TextZoom with the text-focus-loss [4], which we explained in Section 2.

Second, we fine-tune the OCR model on the SR images. Since the SR images are not identical to the high-resolution images due to imperfect super-resolution models, we want to give the OCR model the possibility to learn from these imperfections to improve the recognition accuracy.

We find that splitting the training into these two separate steps improves both the image quality of the SR image and the recognition accuracy compared to adding the loss of the OCR model as an additional term to the first training step.

4.2 Quantization

To reduce model size during inference, we use post-training quantization with different bit widths on specific layers. We use the post-training quantization method Analytical Clipping for Integer Quantization (ACIQ) proposed by [2]. In particular, they propose to clip the tensors using the assumption that parameter values of tensors in neural networks are not uniformly distributed. The values of a tensor often have a bell-shaped curve. Using this fact, they dynamically compute the clipping ranges, reducing the mean square error on a tensor level. They use the assumption that the tensors have a Gaussian or Laplacian distribution [2].

We quantize different layer types and blocks of used models and assessed their performance after quantization. As bit widths, we evaluated 2, 3, 4, 5, 6, 7, and 8 bit.

5 RESULTS

In this section, we present the intermediate scores of the combination of the EDSR, CDistNet, and our two-stage training and show the final results of applying post-training quantization.

EDSR & CDistNet Finetuning. As a first step, we show the text recognition scores on the SR images from EDSR and TBSRN before and after fine-tuning CDistNet to the SR images in table 1. This table shows that fine-tuning CDistNet to the SR images improves the recognition accuracy. Since fine-tuning CDistNet on the high-resolution images yields worse results than fine-tuning on the SR images, we know that the OCR model is learning more than just the characteristics of the dataset we use. This proves that the SR images have certain characteristics that are distinct from the high-resolution images but contain information that helps the OCR model recognize the text. We can also observe that our approach of finetuning outperforms TBSRN and a comination of EDSR and cDist without our finetuning.

SR Model	Finetuned CDistNet	Accuracy			
		Easy	Medium	Hard	Avg.
TBSRN	HR	80.48	64.99	50.63	65.37
TBSRN	SR	84.19	69.81	55.55	69.85
EDSR	HR	79.74	63.22	49.59	64.18
EDSR	SR	**85.05**	**71.37**	**56.81**	**71.08**

Table 1: Effect of fine-tuning CDistNet on on the upsampled images of the super resolution networks. The best score in each column is highlighted bold.

SR Model	Quantization	Bit	Average PSRN	Average Accuracy	Model Size(MB)
TBSRN	None	-	21.15	69.85	2190.08
	SR	4	21.15	70.35	2108.64
	OCR	5	21.15	69.38	1065.89
	SR + OCR	6	21.12	69.85	1031.90
EDSR	None	-	**21.42**	71.08	3457.40
	SR	4	21.23	70.23	2331.32
	OCR	5	**21.42**	70.57	2333.21
	SR + OCR	7	21.41	**71.28**	1329.20

Table 2: Comparison of different quantization settings. We show the best results for quantizing the SR and OCR models individually and together.

Comparison with Interpolation. From an efficiency perspective, we also have to justify using deep learning instead of traditional interpolation techniques. Therefore we compared our approach and the TBSRN with fine-tuned OCR model to bicubic interpolation. The text recognition accuracies for both deep learning approaches are significantly higher than the ones of the OCR model on the interpolated images, with an average accuracy of 71.08 for our approach, 69.85 for TBSRN + fine-tuned OCR model and 61.47 for bicubic interpolation with a fine-tuned OCR model. In addition, the accuracies from table 1 without fine-tuning CDistNet to the SR images are higher than the accuracy on the interpolated images, which shows that the higher accuracy when using deep learning for upscaling does not only come from fine-tuning CDistNet.

Quantization. Finally, we apply the ACIQ quantization to the models. First, we had to select which of the layers we want to quantize. As shown in table 2, different layers react differently to quantization, which results in more or less performance degradation. We found that the following layers in the super-resolution models could be quantized without major performance degradation to 8-bit: convolutional layer, linear layer, multi-head attention layer (TBSRN only), position-wise feed-forward layer, fully connected layers in the STN, and every layer from the head and body (EDSR only). In the OCR model, the following layers could be quantized: linear layer, multi-head attention layer, self-attention layer, and fully connected layer in the localization network.

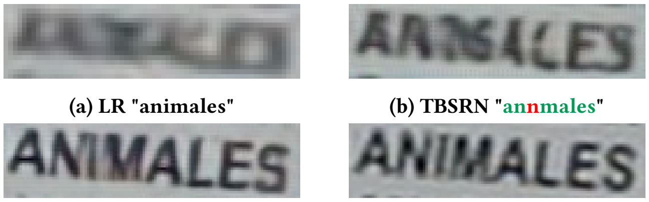
For both super-resolution models, we compare four different cases in table 2: No quantization as a reference, quantizing only the super-resolution model (QSR), quantizing only the OCR model

(QOCR), and quantizing both models together. We tested quantization from 8 down to 2 bits for every configuration and show the results with the smallest possible model relative to the original model size that did not result in major performance degradation in table 2. We sum the corresponding bit size for every quantized parameter and 32 bits for every unquantized parameter to calculate the theoretical model sizes.

We can observe that the text recognition accuracy is more sensitive to quantization with variations of over 1% when using the EDSR and almost 0.97% when using TBSRN, while the maximal PSNR variation lies at only 0.19 for the EDSR and only 0.03 for TBSRN.

Both the mean PSNR and the mean accuracy, as well as every PSNR and accuracy on the individual test subsets of our 7-bit quantized fine-tuned model combination of the EDSR and CDistNet perform better as the unquantized combination of TBSRN with CDistNet while being over 800 MB smaller in theoretical model size. The mean accuracy of our 7-bit quantized model is even higher than of our unquantized model. This can be explained with possible regularization through quantization and can also be seen in the comparison of the mean accuracy of the 4-bit quantized TBSRN and unquantized CDistNet of 70.35 with the mean accuracy of the unquantized TBSRN and unquantized CDistNet of 69.85.

Two examples from the hard test dataset where our approach performs better than TSBRN even with fine-tuned CDistNet can be seen in Figure 2. We show the low-resolution and high-resolution images annotated with the correct text label together with the TBSRN SR image and our fine-tuned EDSR SR image, which are annotated with the CDistNet text predictions on them. Correctly predicted letters are coloured green, and incorrectly predicted letters are coloured red.



(a) LR "animales"

(b) TBSRN "annmales"

(c) Ours "animales"

(d) HR "animales"

Figure 2: Comparison of TBSRN + CDistNet quantized to 6 bits and EDSR + CDistNet quantized to 7 bits on two example test images.

It has to be mentioned that the overall smallest model is achieved when quantizing TBSRN together with CDistNet.

Since we achieve higher scores with a comparably small model, which is underlined by the noticeably improved readability shown in Figure 2, we argue that using our model offers a better trade-off between model size and prediction results than the quantized TBSRN with CDistNet and definitely is a better trade-off than the original TBSRN with CDistNet.

6 CONCLUSION AND FUTURE WORK

We demonstrated how text super-resolution improves OCR even with a very powerful state-of-the-art model like CDistNet. We further showed that for our use case, text recognition aware training is sufficient to optimize an image super-resolution model to work on text images, and using a general-purpose super-resolution yields better results compared to a specially designed text super-resolution model due to the possibility of knowledge transfer from non-text related super-resolution datasets.

Furthermore, we achieved state-of-the-art performance at a smaller theoretical model size than the previously best model on this dataset. This would lead to less memory and energy consumption, in addition to faster execution times on low-power devices that can take advantage of this low precision representation.

REFERENCES

[1] Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 126–135.

[2] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. 2019. Post-training 4-bit quantization of convolution networks for rapid-deployment. arXiv:1810.05723 [cs.CV]

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. (2012).

[4] Jingye Chen, Bin Li, and Xiangyang Xue. 2021. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12026–12035.

[5] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. 2020. Text Recognition in the Wild: A Survey. arXiv:2005.03492 [cs.CV]

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*. Springer, 184–199.

[7] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A Survey of Quantization Methods for Efficient Neural Network Inference. arXiv:2103.13630 [cs.CV]

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[9] Yue He, Chen Chen, Jing Zhang, Juhua Liu, Fengxiang He, Chaoyue Wang, and Bo Du. 2021. Visual Semantics Allow for Textual Reasoning Better in Scene Text Recognition. arXiv:2112.12916 [cs.CV]

[10] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5197–5206.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).

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

[13] Wanjie Sun and Zhenzhong Chen. 2020. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing* 29 (2020), 4027–4040.

[14] Hanh TM Tran and Tien Ho-Phuoc. 2019. Deep laplacian pyramid network for text images super-resolution. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 1–6.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[16] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. 2020. Scene text image super-resolution in the wild. In *European Conference on Computer Vision*. Springer, 650–666.

[17] Roman Zeyde, Michael Elad, and Matan Protter. 2010. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*. Springer, 711–730.

[18] Tianlun Zheng, Zhineng Chen, Shancheng Fang, Hongtao Xie, and Yu-Gang Jiang. 2021. CDistNet: Perceiving Multi-Domain Character Distance for Robust Text Recognition. arXiv:2111.11011 [cs.CV]