Compression of SRGAN

Aditya Iswara, Arvind Subramanian, Raina Kolluri

{iswara, arvindvs, rainak}@stanford.edu

Introduction

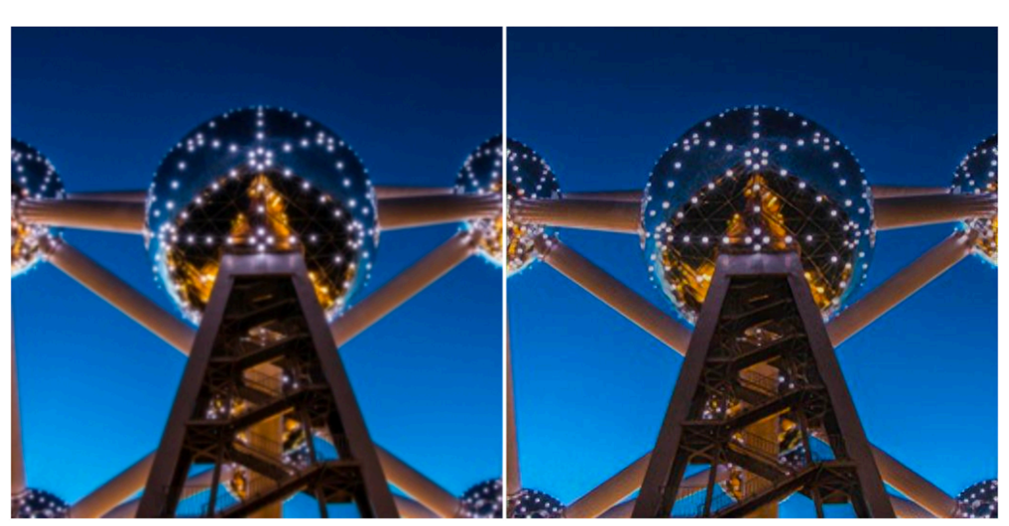
Recently, the applications of deep neural networks have becoming increasingly prevalent. An example of this is SRGAN [1], which is used for the task of superresolution: creating a high-resolution counterpart of a low-resolution image [2]. This is problematic when these models need to be used in situations where the network must perform inference quickly and fit within a certain amount of memory. An example of this is deploying superresolution on a mobile device, which would greatly improve the resolution of photos on the device but requires a reduction in the size of the model. Neural network compression is the task of minimizing the size of a network while maximizing its accuracy [3]. In our project, we apply compression techniques such as pruning and knowledge distillation on SRGAN. We seek to maximize performance while also reducing size.

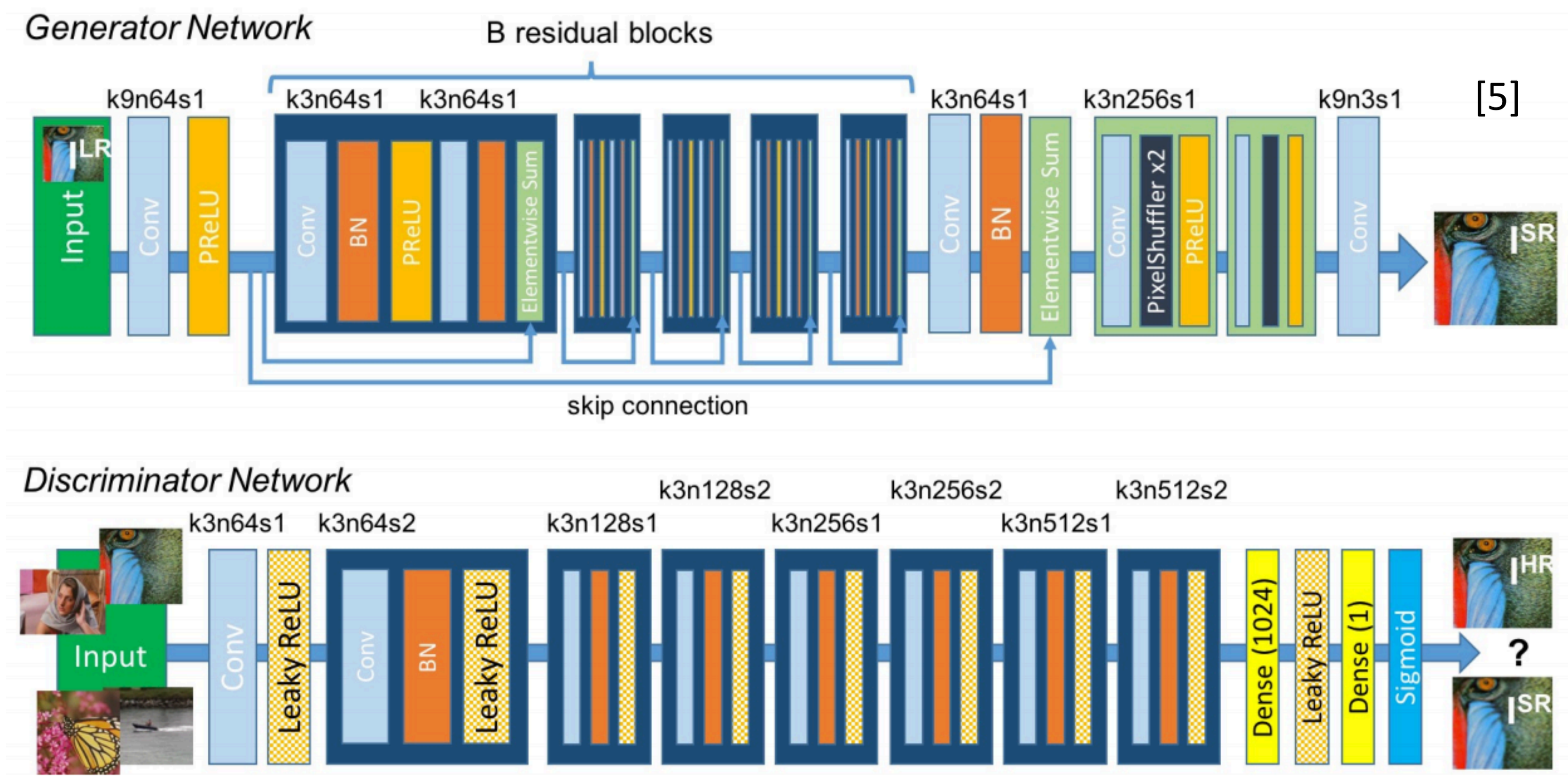
Dataset

Div2K

1000 2K resolution images with corresponding low-resolution images for 2, 3, and 4 downscaling factors

- 800 Training
- 100 Validation
- 100 Testing





SRGAN

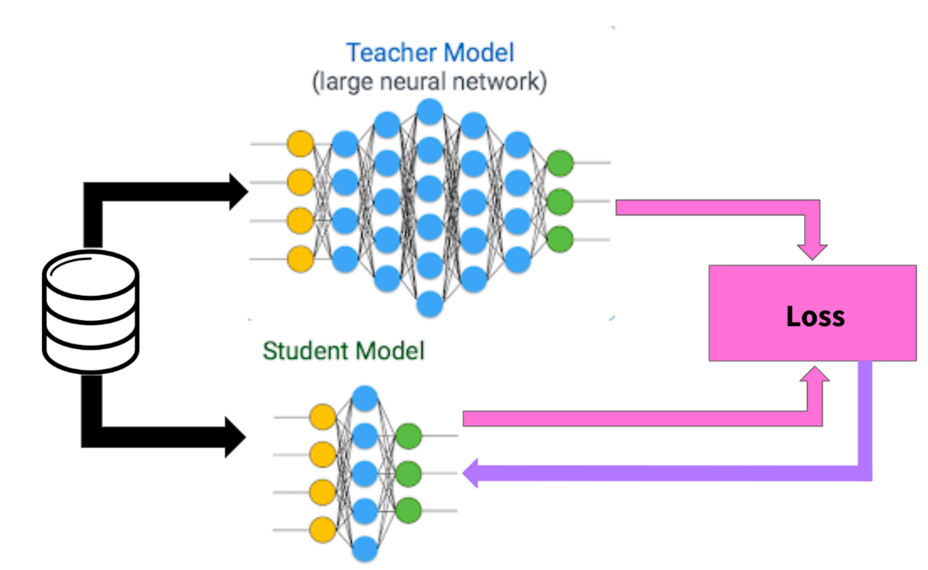
- Model Size: ~40MB [4]
- Our Compression Techniques focused on reducing the number of Residual Blocks in the Original Model (B = 16)
- Goal is to Reduce Model Size to ~20MB

Pruning

- This technique simply involves removing residual blocks from the generator network and retraining the entire model from scratch to check performance
- Removed 10 residual blocks (B = 6) to get model size of ~20MB

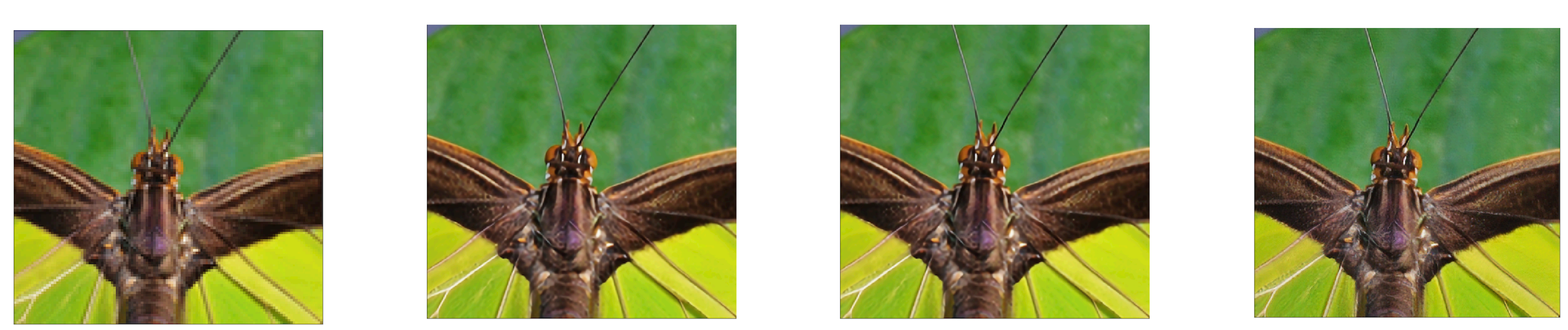
Knowledge Distillation

- Takes logits layer immediately before softmax layer and trains our compressed model using modified loss function
- Original model acts as a "teacher" and the new, smaller model (same model as pruning) is the "student"
- Knowledge distillation only trains student generator



Results

- SSIM Scores
  - Pruning: .3455
  - Knowledge Distillation: .5011
  - Original Model: .6688 (Trained on Different Dataset)



Downscaled Image | Pruned Model | Knowledge Distillation | SRGAN

Discussion

- Minimal visual distinction between knowledge distillation and pruning
- SSIM scores do show improvements by using KD
- Both techniques performed as expected, although to a lower degree
- Still needs improvement to reach performance close to the original SRGAN

Shortcomings

- Pruning was done without intelligently selecting which weights contribute least to the generator output
- Both techniques use the same model architecture, which may not be optimal

Future Work

- "Smarter" Pruning
  - Looking at where the weights are close to 0 in training and removing those layers from the network
- Varied KD architecture
  - Use different layer sizes to find optimal performing student generator
- Extensive hyperparameter tuning

References

[1] Agustsson, Eirikur, and Radu Timofte. "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study." Google AI Blog, Google Research, 24 Aug. 2017, https://ieeexplore.ieee.org/document/8014884
}

[2] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

[3] Cheng, Y., Wang, D., Zhou, P., Zhang, T. "A Survey of Model Compression and Acceleration for Deep Neural Networks." IEEE Signal Processing Magazine, 8 Sept. 2019.

[4] https://github.com/krasserm/super-resolution

[5] Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017.