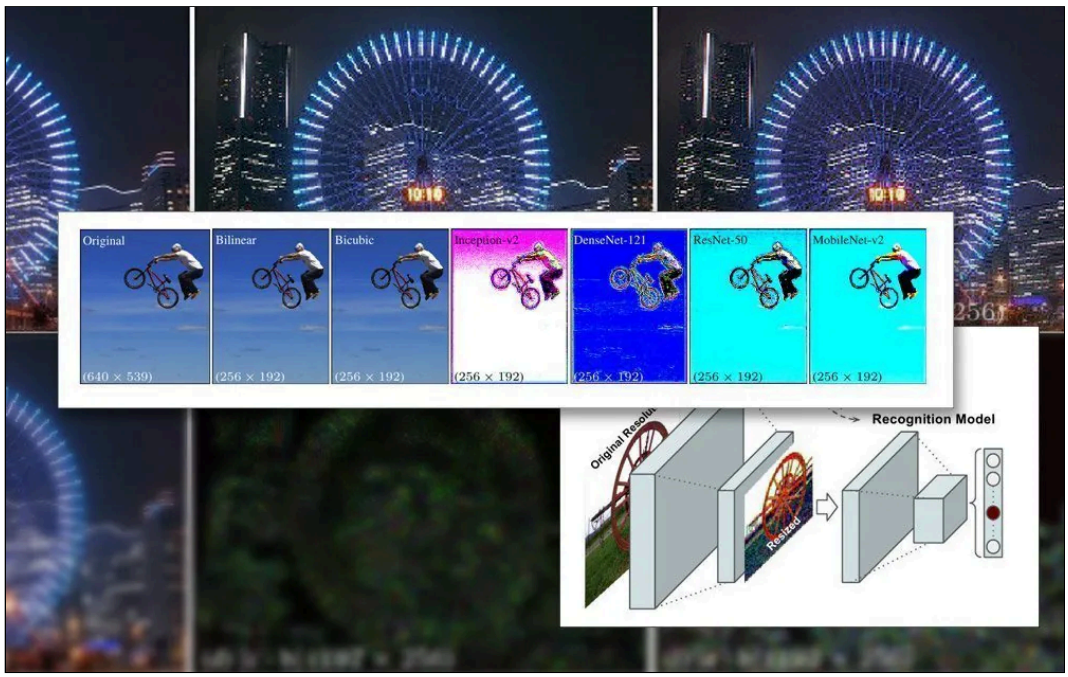
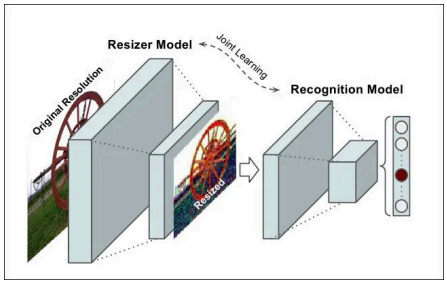
Better Machine Learning Performance Through CNN-based Image Resizing

Originally published at https://www.unite.ai/better-machine-learning-performance-through-cnn-based-image-resizing/



Google Research has proposed a new method to improve the efficiency and accuracy of image-based computer vision training workflows by improving the way that the images in a data set are shrunk at the pre-processing stage.

In the paper *Learning to Resize Images for Computer Vision Tasks*, researchers Hossein Talebi and Peyman Milanfar utilize a CNN to create a new hybrid image resizing architecture that produces a notable improvement in recognition results obtained over four popular computer vision datasets.



The proposed joint framework for recognition and resizing. Source: https://arxiv.org/pdf/2103.09950.pdf

The paper observes that the rescaling/resizing methods that are currently used in automated machine learning pipelines are decades out of date, and frequently use just basic bilinear, bicubic and nearest neighbor resizing – methods which treat all pixels indiscriminately.

By contrast, the proposed method augments the image data via a CNN and incorporates that input into the resized images that will ultimately pass through the model's architecture.

Image Constraints in AI Training

In order to train a model that deals with images, a machine learning framework will include a pre-processing stage, where a disparate variety of images of various sizes, color spaces and resolutions (that will contribute to the training dataset) are systematically cropped and resized into consistent dimensions and a stable, single format.

In general this will involve some compromise based around the PNG format, where a trade-off between processing time/resources, file size and image quality will be established.

In most cases, the final dimensions of the processed image are very small. Below we see an example of the 80×80 resolution image at which some of the earliest deepfakes datasets were generated:

Since faces (and other possible subjects) rarely fit into the required square ratio, black bars may need to be added (or wasted space permitted) in order to homogenize the images, further cutting down the actual usable image data:

Here the face has been extracted from a larger image area until it is cropped as economically as it can be in order to include the entire face area. However, as seen on the right, a great deal of the remaining area will not be used during training, adding greater weight to the importance of the image quality of the resized data.

As GPU capabilities have improved in recent years, with the new generation of NVIDIA cards outfitted with increasing amounts of video-RAM (VRAM), average contributing image sizes are beginning to increase, though 224×224 pixels is still pretty standard (for instance, it is the size of the ResNet-50 dataset).

An unresized 224×244 pixels image.

Fitting Batches Into VRAM

The reason the images must all be the same size is that gradient descent, the method by which the model improves over time, requires uniform training data.

The reason the images have to be so small is that they must be loaded (fully decompressed) into VRAM during training in small batches, usually between 6-24 images per batch. Too few images per batch, and there is not enough group material to generalize well, in addition to extending the training time; too many, and the model may fail to obtain necessary characteristics and detail (see below).

This 'live loading' section of the training architecture is called the latent space. This is where features are repeatedly extracted from the same data (i.e. the same images) until the model has converged to a state where it has all the generalized knowledge it needs to perform transformations on later, unseen data of a similar type.

This process generally takes days, though it can take even a month or more of constant and unyielding high volume 24/7 cogitation to achieve useful generalization. Increases in VRAM size are only helpful up to a point, since even minor increments in image resolution can have an order-of-magnitude effect on processing capacity, and related effects on accuracy that may not always be favorable.

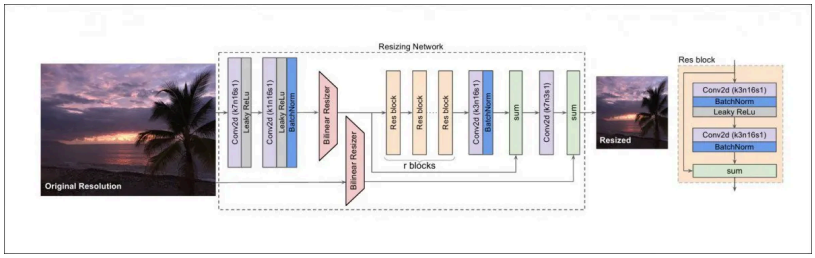
Using greater VRAM capacity to accommodate higher batch sizes is also a mixed blessing, as the greater training speeds obtained by this are likely to be offset by less precise results.

Therefore, since the training architecture is so constrained, anything that can effect an improvement within the existing limitations of the pipeline is a notable achievement.

How Superior Downsizing Helps

The ultimate quality of an image that will be included in a training dataset has been proven to have an improving effect on the outcome of training, particularly in object recognition tasks. In 2018 researchers from the Max Planck Institute for Intelligent Systems contended that the choice of resampling method notably impacts training performance and outcomes.

Additionally, prior work from Google (co-written by the new paper's authors) has found that classification accuracy can be improved by maintaining control over compression artefacts in dataset images.



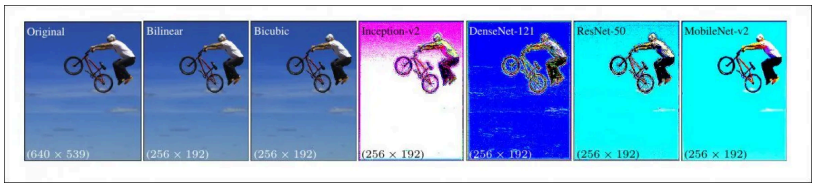
The CNN architecture for the Google Research proposed downsampling algorithm.

The CNN model built into the new resampler combines bilinear resizing with a 'skip connection' feature that can incorporate output from the trained network into the resized image.

Unlike a typical encoder/decoder architecture, the new proposal can act not only as a feed-forward bottleneck, but also as an inverse bottleneck for up-scaling to any target size and/or aspect ratio. Additionally, the 'standard' resampling method can be swapped out for any other suitable traditional method, such as Lanczos.

High Frequency Details

The new method produces images that in effect seem to 'bake' key features (that will ultimately be recognized by the training process) directly into the source image. In aesthetic terms, the results are unconventional:



The new method applied across four networks – Inception V2; DenseNet-121; ResNet-50; and MobileNet-V2. The results of the Google Research image downsan images with obvious pixel aggregation, anticipating the key features that will be discerned during the training process.

The researchers note that these initial experiments are exclusively optimized for image recognition tasks, and that in tests their CNN-powered 'learned resizer' was able to achieve improved error rates in such tasks. The researchers intend in future to apply the method to other types of image-based computer vision applications.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More