Attention Is All You Need

We looked into how the depth of a convolutional neural network impacts accuracy when it comes to recognizing large-scale images.

ConvNets have been successful in the large-scale image and video recognition due to ImageNet and high-performance computing systems. This paper addresses the depth of ConvNet architectures by adding more convolutional layers and releasing two best-performing models to facilitate further research.

The design of ConvNet layers follows the same standards established by Ciresan et al. (2011) and Krizhevsky et al. (2012). Our ConvNet configurations are different from the ones used in the top-performing entries of the ILSVRC-2012 and 2013 competitions, with smaller conv. layer widths and receptive fields.

The ConvNet training procedure follows Krizhevsky et al. (2012) and requires fewer epochs due to implicit regularisation and pre-initialization of certain layers. We used random initialization and training image rescaling to obtain fixed-size 224×224 ConvNet input images, which were randomly cropped from rescaled training images and underwent random horizontal flipping and RGB color shift. The fully-convolutional network is applied densely over the rescaled test image, resulting in a class score map with a variable spatial resolution. Multi-crop evaluation is complementary to dense evaluation and provides a 3.75x speedup on an off-the-shelf 4-GPU system.

To analyse the identify the performance of ConvNet models, we tested them at a single scale and used the layer configurations mentioned in Sect. 2.2.2 We found that local response normalization does not improve performance, but that the classification error decreases with increased depth. Scale jittering at training time leads to significantly better results than training on images with the fixed smallest side, confirming that training set augmentation by scale jittering is helpful for multi-scale image statistics.

Multi-crop evaluation is beneficial to dense evaluation and offers a 3.75x faster performance when using a standard 4-GPU system. Our highly developed ConvNets surpass the performance of existing models which were able to achieve success in ILSVRC-2012 & 2013 competitions.

Deep convolutional networks can achieve state-of-the-art performance on ImageNet challenge datasets with increased depth

We are extremely thankful to the ERC grant VisRec no. 228180 for their support which enabled this work. Furthermore, we would also like to express our gratitude to NVIDIA Corporation for donating the GPUs utilized in this research.