OPTIMIZATION OF TENSORFLOW FOR CPU IN

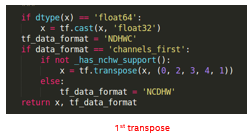
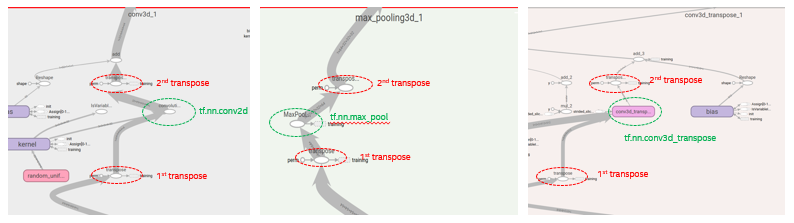
IMAGE SEGMENTATION WITH 3D U-NET CNN WITH KERAS

BACKGROUND ON KERAS

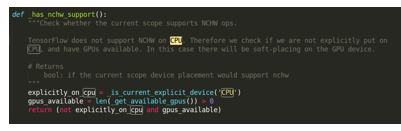
Keras is a high-level API for use with deep learning models, it is user-friendly, modular, and easily expandable. It makes model building easy and connects everything together.

Our engineers found that when using "NCHW" format dataset Keras would insert two additional "Transpose" operations before and after every CNN operation, impacting every Keras layer, including conv2d, conv3d, batchnorm, batchnorm3d, maxpooling, and maxpooling3d.

According to Keras. Tensorflow causes this condition by supporting two data formats; Channel\_First (NCDHW) and Channel\_Last (NDHWC), and while GPUs support both formats, CPU’s support only Channel\_Last, making Keras perform a transpose when data input format is Channel\_First.



**Fig 3.** Author: Zhu, Wenxi. Source: “OOB Performance – 3D Unet”



This condition creates a significant overhead in typical CNN models that in this case contribute to roughly 34% of total inference time in 3D-UNet. We initiated a PR with Google aimed at removing this overhead, suggesting the following steps:

1. Detect the "transpose + conv2d + transpose" pattern;
2. Delete those 2 "transpose" ops, and make "conv2d"'s format "NCHW", and
3. Replace "conv2d" with "\_MklConv2D", which supports both "NCHW" and "NHWC" format.

The dialogue with Google can be located at:  <https://github.com/tensorflow/tensorflow/pull/23152>