README

This repository contains the PyTorch re-implementation of the sparsely-gated MoE layer described in the paper [Outrageously Large Neural Networks](https://arxiv.org/abs/1701.06538) ( [arXiv:1701.06538](https://arxiv.org/abs/1701.06538) [cs.LG]), linked by the paper [[2209.01667] A Review of Sparse Expert Models in Deep Learning](https://arxiv.org/abs/2209.01667) ([arXiv:2209.01667](https://arxiv.org/abs/2209.01667)[cs.LG]).

The file PROGETTO\_NN\_MoE.ipynb illustrates how to build, train and evaluate the MoE model with inputs and outputs taken from CIFAR10. To run it:

jupyter nbconvert --execute PROGETTO\_NN\_MoE.ipynb

The code is based on the TensorFlow implementation that can be found here:

[tensor2tensor/tensor2tensor/utils/expert\_utils.py at master · tensorflow/tensor2tensor · GitHub](https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/utils/expert_utils.py) .