I started experimenting with Transformers with the v3 data. jrb20

was my first transformer model. It’s just avanilla 4 layer transformer that takes embeddings of the 1050 features as a sequence and the model just has a single linear neuron at the end on the concatenated sequence output from the transformer. My newer models are bigger mixture of experts of small transformer like models and the routing model is a hypernetwork. These things are much trickier to train, but the results seem quite promising. I believe the hard routing (and the softmax attention in the transformers) give these models GBDT like properties, with the flexibility of NNs (pre-training, better loss functions, better optimizers, architectural tricks, etc).