**Butterfly encoder-decoder linear layer replacement – experiments**

The goal of this project is to reduce the number of parameters built into the Vision Transformer image classifier by means of replacing its linear layers on a 1-to-1 basis with a SELL (structured efficient linear layer) involving a butterfly network (library cloned from <https://github.com/HazyResearch/butterfly> ) projecting input to an intermediate dimension, followed by a square linear layer for linear transformation within that dimension, followed by a second butterfly network. Assuming a linear layer of shape (input dimension m, output dimension n) and intermediate dimension of size t, our replacement layer consisted of a butterfly network of shape followed by a linear layer of shape , followed by a second butterfly network of shape . The intermediate linear layers within these SELLs are initialized based on the original linear layer by means of multiplication (from either side) of the given linear layer matrix with the transposes of the matrices describing the linear transformations performed by the butterfly networks (which are sampled from the Johnson-Lindenstrauss matrix distribution). Training optimization was performed with the AdamW algorithm, just as described in the ViT paper (although with a batch size of 16 instead of 2048 due to memory constraints, and a fixed learning rate of 0.00004).

Various experiments were performed on a subset of ImageNet called ImageNette (a subset consisting of 10 out of the 200 classes included in ImageNet), beginning with

1. Application of the pretrained Vision Transformer obtained from the Timm library, achieving excellent classification performance (near 100% precision)
2. Replacement of all linear layers with the aforementioned SELL with for code verification purposes, achieving excellent classification performance (near 100% precision). It is noteworthy that this step actually increases the number of parameters in the model and is overfit to the ImageNette training set (and so lower performance may occur on the ImageNet training set), this step is only to eliminate the possibility of bugs in the code.
3. Replacement of the linear layers with the aforementioned SELL with to lower the number of parameters in use in the model. Both experiments yielded very poor precisions, with precisions in the range of 15% or lower.
4. Replacement of the linear layers with the aforementioned SELL with to lower the number of parameters in use in the model by ~3 million (out of ~86 million parameters total in the original model). Only the “qkv” and “projection” linear layers of the attention blocks were replaced\*, leaving the MLP linear layers as-is (although the entire model was trained together, with these MLP layers being fine-tuned to work properly with the outputs of the new SELL replacements). This experiment too yielded poor weighted average precision of 57%, with precisions ranging from 37%-77%. The author predicts that further training may further increase precision somewhat.

\*The reasoning behind this experiment was due to author’s intuition that the butterfly layer works best as a feature extractor due to its prevalence in the Fourier, Hadamard, and Cosine literature, and that the actual processing of these features must be performed with a full-sized MLP with regular linear layers.

Unfortunately the results from this series of experiments were rather disappointing, the author hopes further research into this field may yield better results.