1 Dataset

This is a multi-label dataset:

- Each input is a 480 length vector representing 60 8-dimensional vectors. Each 8-dimensional vector is an "attribute".
- Each label is a 60 length vector representing the target value of each of the 60 "attributes". There are 8 target values.
- Each "attribute" can take 8 values (the 8 target values) represented by a random vector from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Noise is added for each sample.

2 Model

A simple 4 layer MLP with:

- Layernorm after the 4 layers. This is fed into a classifier to predict the 60 labels.
- A single residual stream from the input to the layer norm.
- GeLU activation function.
- 512-512-256-512 hidden layer architecture

3 Steering adaptor

A forward hook is inserted at the 256-dim layer to extract activations for the contrastive pairs. The inputs for generating the contrastive pairs are made as

- Selecting one attribute to target steering.
- Positive examples set this attribute to 0.
- Negative examples set this attribute to 1-7.

CAA simply takes the difference of means of these two activations as a steering vector. A forward hook is registered at the same 256-dim layer which simply adds the scaled (parameterised by λ) steering vector to the output and returns it for the next layer.

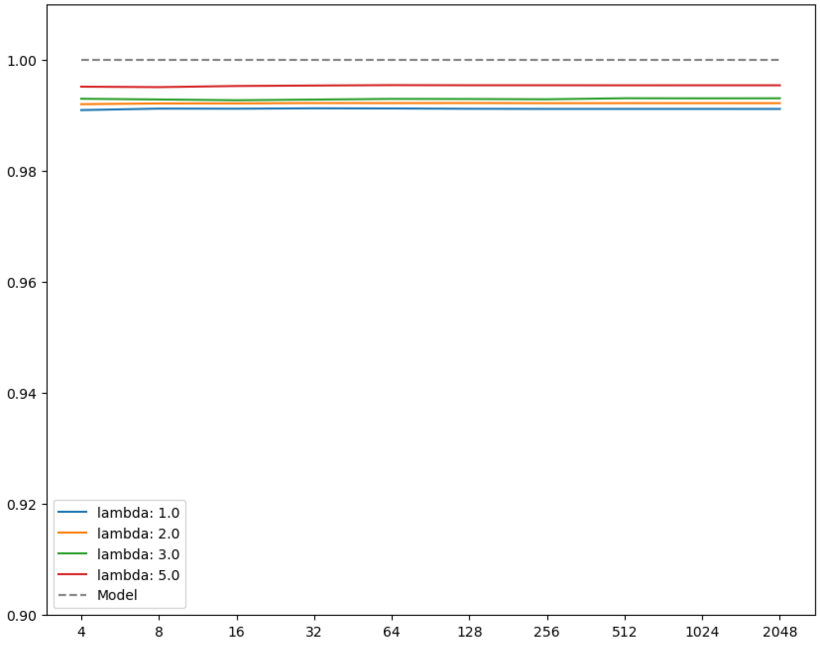
4 Steering metric

On a held out set of negative examples the model (with the trained steering adaptor) is evaluated. The target is the positive labels, meaning successfully steering the selected attribute without effecting the other attributes.

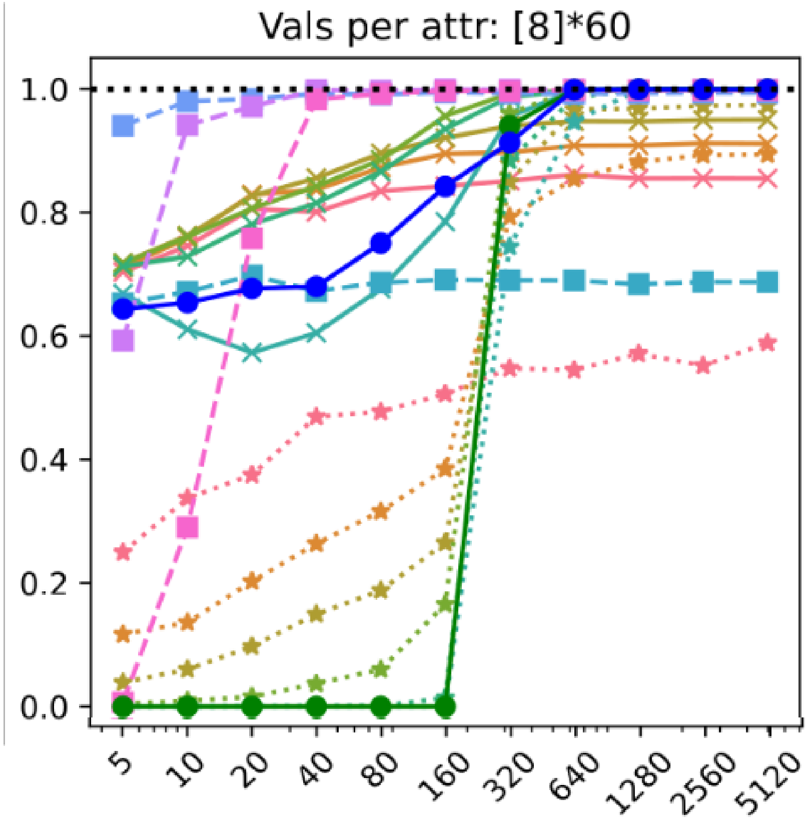
The accuracy of the steered model on this held out test set is reported as the steering metric. A higher score means better steering.

5 Current results

As shown in Figures 1a and 1b there is clearly something wrong with the implementation.



(a) Current results following the paper setup and Dima's email.



(b) Target graph. The square marked lines are the different values of lambda. Light blue -> purple is increaing values of lambda.

6 Ablations

To narrow down where the potential issue is I have carried out the following tests and found the following results

- Running with and without the adaptor: There is a significant difference in the output on average suggesting that the adaptor is steering the model.
- Increasing the size of lambda: This results in a larger impact on the outputs but the curves of the paper are still not present.
- Running the steered model on the negative inputs used in training: Interestingly the model does not steer the output towards the positive labels even though this is the test data. I feel this is likely the issue.