Steering Clear Reproduction: CAA

Skye Purchase

20 June 2025

I follow the original paper [**steering-clear**] in all regards and any additional assumptions I make are explained below. The paper describes the overview of how experiments were run but specific details are still missing allowing some room for interpretation.

I find that with a range assumptions across a number of trials that I am unable to fully reproduce the CAA plots the paper presents. The primary issue I find is with the steerability metric stated in the paper, of total accuracy, where just the steered attribute accuracy results in a plot closer to the paper.

1 Dataset

This is a multi-label dataset:

- Each input is a 120 length vector representing 60 2-dimensional vectors. Each 2-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.

I use only 2-dimensional vectors rather than 8 for reduced computation costs and the fact that the new paper uses this instead. Furthermore, I only use $500,000$ examples rather than $2,000,000$ due to memory constraints but find that this does not affect model accuracy. The paper does not state the exact method of generating anchor points (3) but achieving 100% on the model should suffice. I use a standard deviation of 0.01 when adding noise to the samples thus insuring the datapoints are separable.

2 Model

A simple 4 layer MLP with:

- Layernorm after the 4 layers. This is fed into a classifier to predict the 60 labels.
- GeLU activation function.
- 512-512-256-512 hidden layer architecture.
- A 60 head classifier.

I test 3 types of residual streams:

- A single stream from input to layernorm.
- A residual over every single layer.
- No residual streams anywhere.

Figures 1a, 1b and 1c show the train curves for the different models. All of these eventually reach near 100% accuracy. The best model is chosen based on best validation accuracy and is chosen on the first instance of the best validation loss.

Figures 2a, 2b and 2c show the accuracy of the model across all attributes and their values. All other values are filled with a uniform random attribute. All fills are presented to show that there is no bias towards a particular fill value.

From these plots I am fairly confident that my setup for the dataset and the training of the model is sound though there are clearly differences between how the residual streams are applied.

3 Steering adaptor

A forward hook is inserted at the 256-dim layer to extract activations for the contrastive pairs. In the case of the residual stream per layer model the hook is inserted after the residual stream.

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.

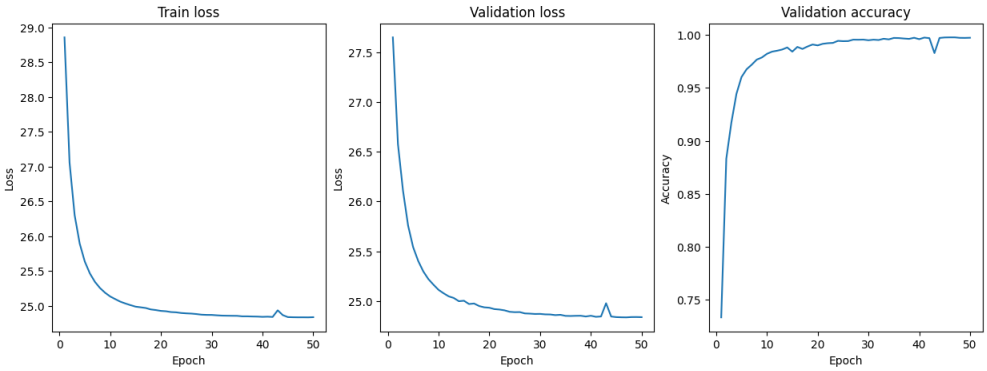
To get repeated runs, a set of attributes is chosen[1] and the above process is applied to each. In the end I run 20 repeats.

Figure 3 shows the cosine similarity of a sample of attributes. The similarity between pairs is very high as expected as the other 59 attributes are identical. The cross similarity is much lower showing that there is a variety of examples that are present when training the adaptor. This plot is essentially the same across the models with minor differences in the exact similarity values.

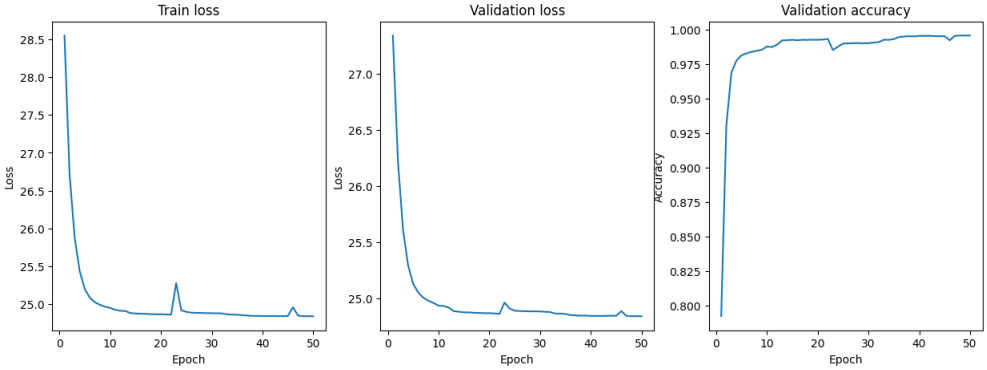
Figures 4, 5 demonstrate the effect of the CAA steering on the different attribute values for the target attribute. The remaining attribute values are randomly selected. The experiment is run over a range of example values for each 20 repeats and for each experiment the mean over 100 test inputs is returned.

Clearly demonstrated is that 4 and 8 examples do not achieve the same efficacy as the other examples regardless of the strength of the steering vector. This does not take into account the effect on the attributes nor the softmax prob of the other values. These experiments demonstrate that the steering vectors are able to steer effectively.

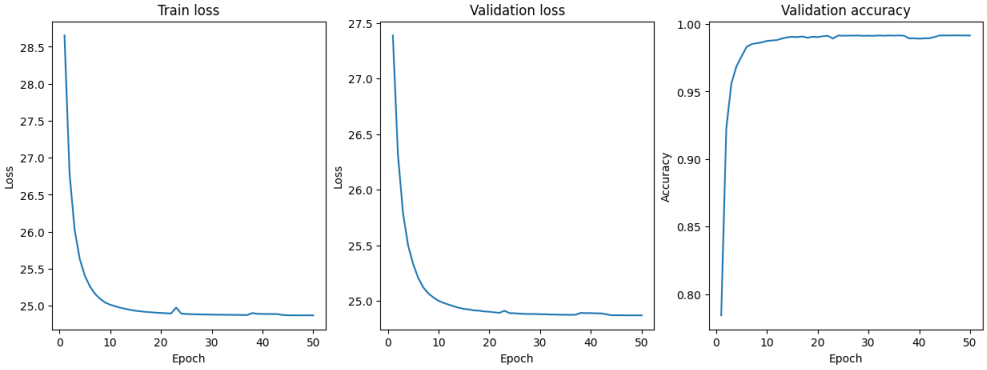
[1] For ease of implementation these are just the first n attributes



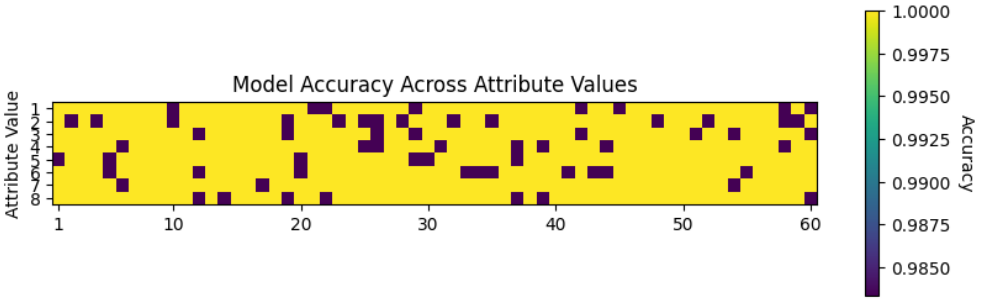
(a) Train and loss curves for the MLP without any residual streams.



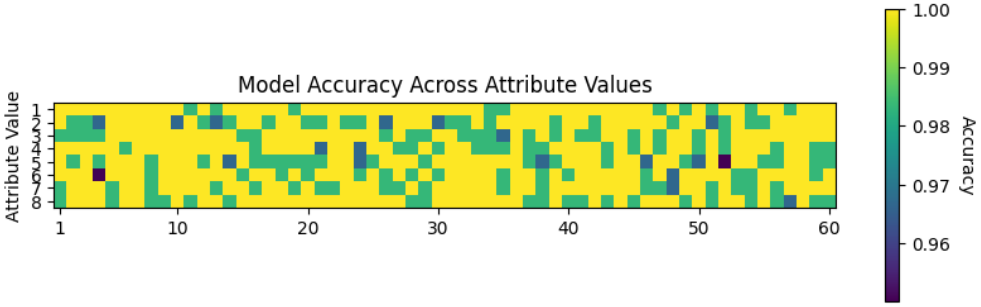
(b) Train and loss curves for the MLP with a residual stream per layer.



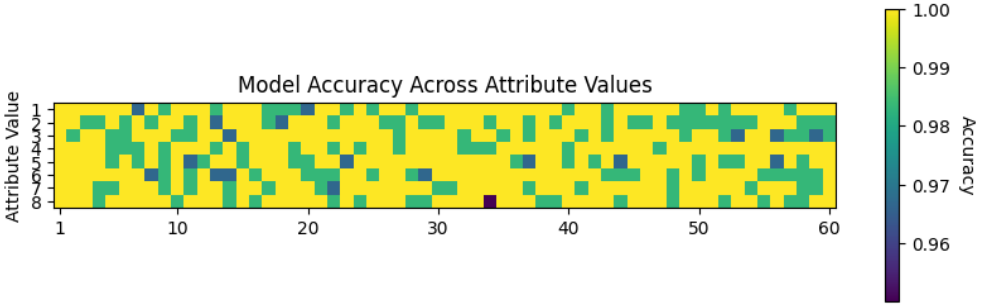
(c) Train and loss curves for the MLP with a single residual stream from input to layer-norm.



(a) The accuracy of the non-residual model on each attribute value.



(b) The accuracy of the full residual model on each attribute value.



(c) The accuracy of the single residual model on each attribute value.

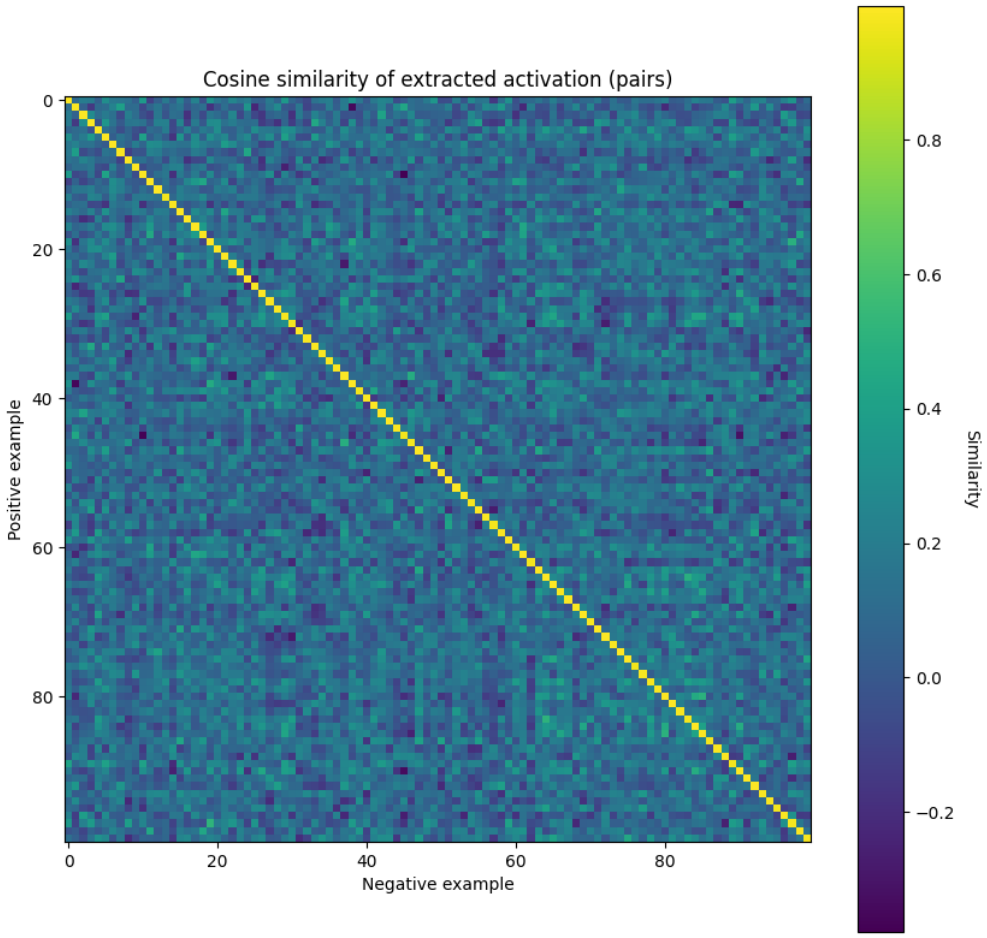


Figure 3: The cosine similarity between a sample of positive and negative pairs. This is for a specific attribute.

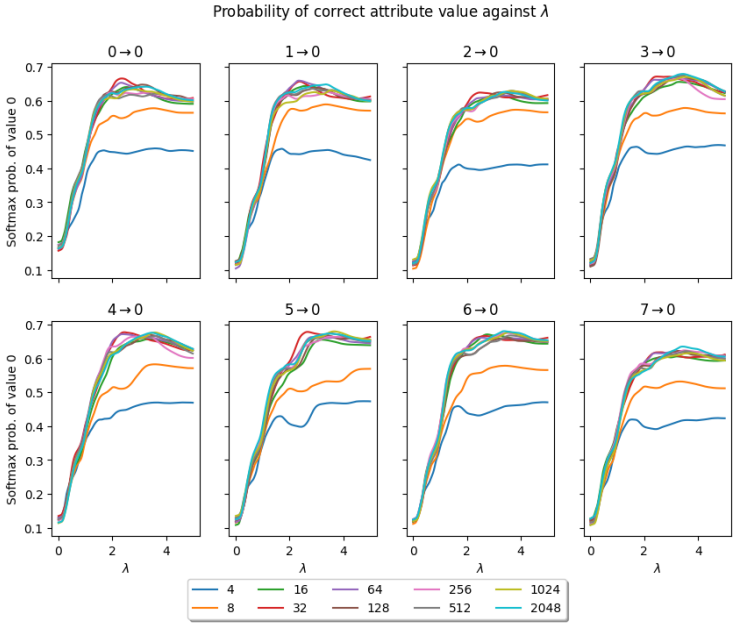


Figure 4: The softmax probability of the target label (0) given the input label as a function of the scaling parameter λ. This is the model without any residual streams.

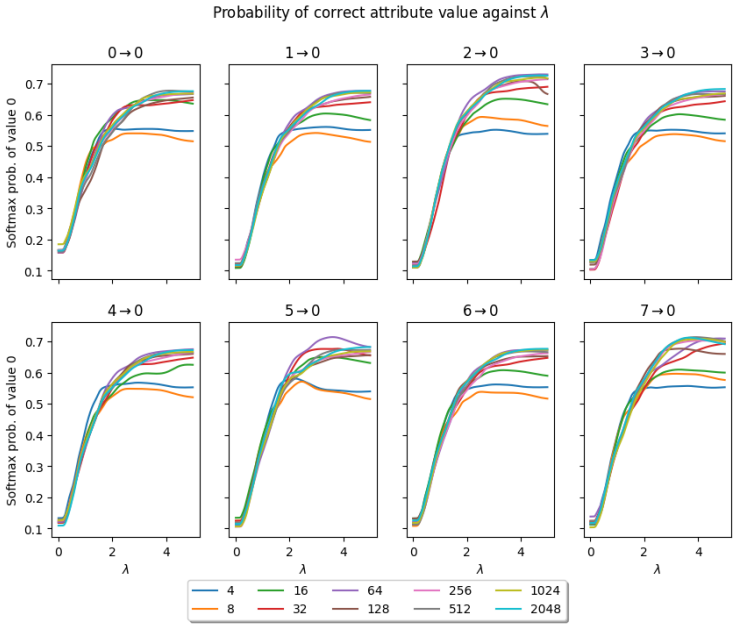


Figure 5: The softmax probability of the target label (0) given the input label as a function of the scaling parameter λ. This is the model with a residual streams per layer.

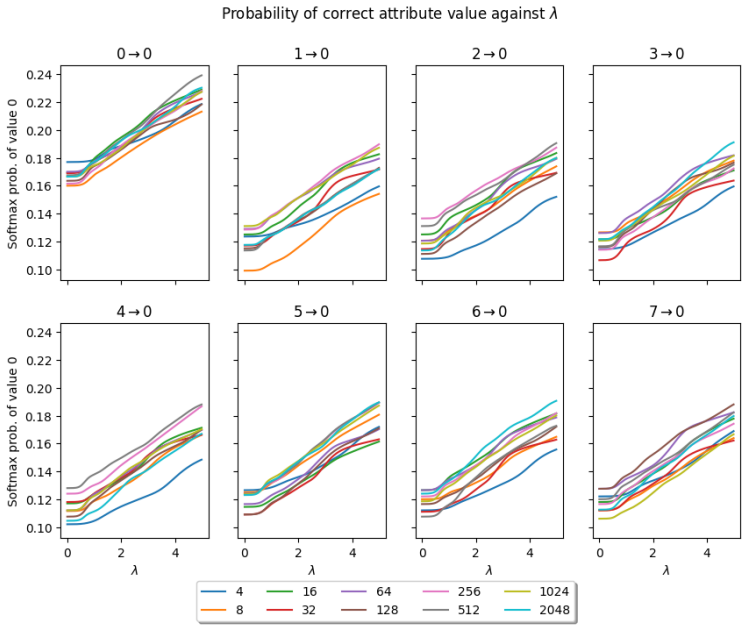


Figure 6: The softmax probability of the target label (0) given the input label as a function of the scaling parameter λ. This is the model with a single residual stream from input to layernorm.

4 Steering metric

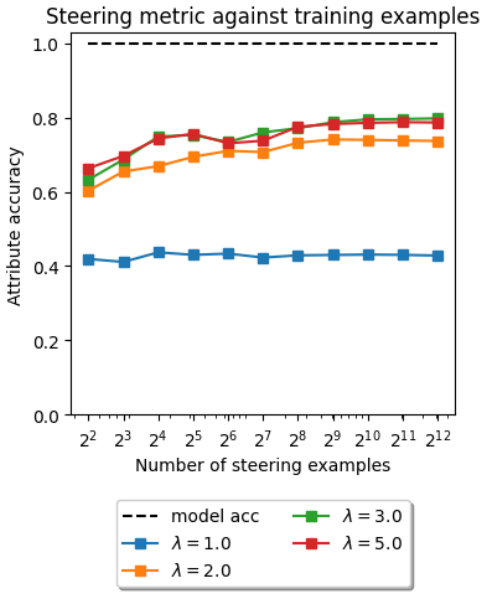
A subset of 1000 of the contrastive pairs above are withheld during training. The negative generating inputs are fed through the model with the trained steering adaptor and the output recorded. The goal is for the output to match the positive generating labels.

There are two metrics that I test:

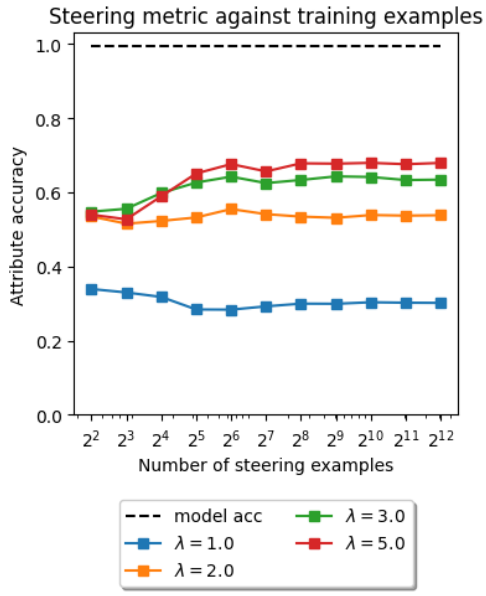
- The accuracy on the steered attribute alone.
- The accuracy on all the attributes (this is the one stated in the paper).

Figures 7a, 7b and 7c show the attempt at reproducing the paper figures for the top-left plot in Figure 1. This only focuses on the CAA approach. This uses a different metric to the paper focusing only on attribute accuracy rather than total accuracy and the trends are similar to those in the paper. However, these plots are still not the same as the papers plots especially as they do not use the same metric.

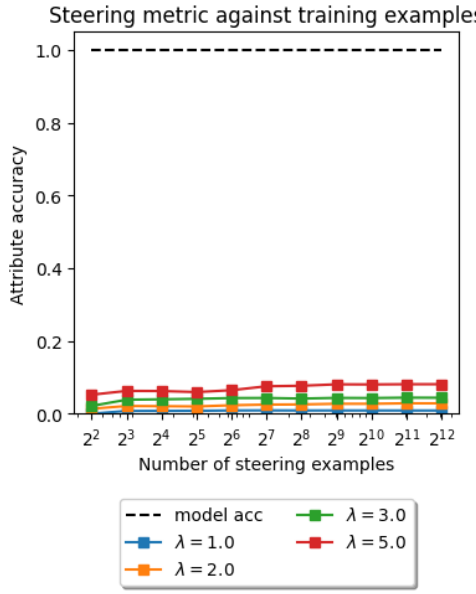
Figures 8a, 8b and 8c show the reproduction using the stated steering metric of total accuracy across all target labels.



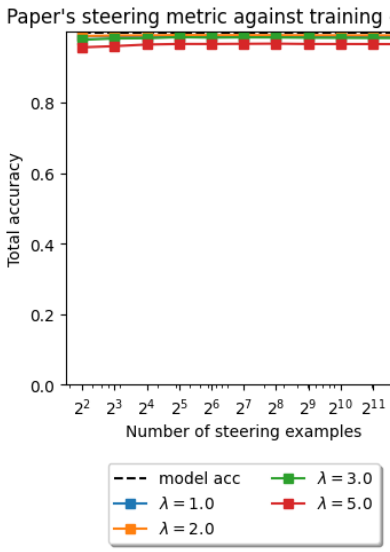
(a) Standard model steering accuracy on the steered attribute alone.



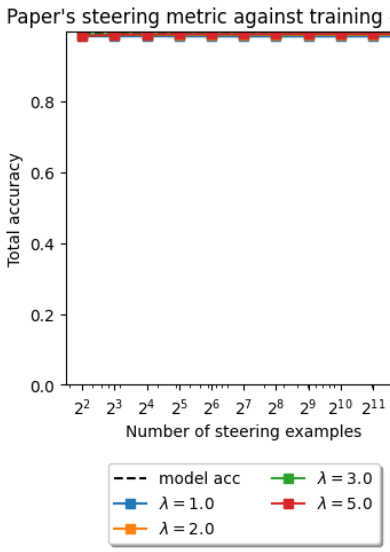
(b) Residual model steering accuracy on the steered attribute alone.



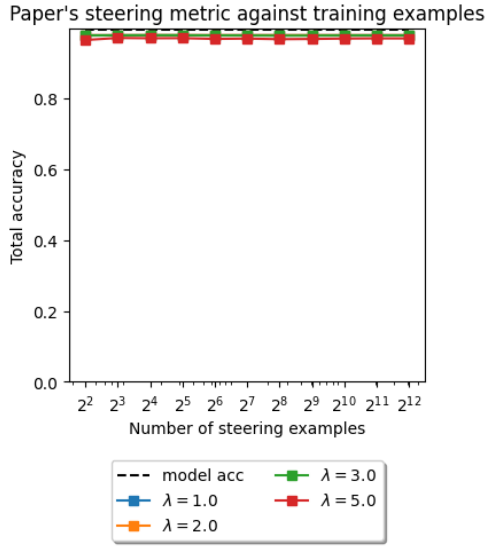
(c) Single residual stream model steering accuracy on the steered attribute alone.



(a) Standard model steering accuracy on the steered attribute alone.



(b) Residual model steering accuracy on the steered attribute alone.



(c) Single residual stream model steering accuracy on the steered attribute alone.