A. Appendix

A.1. Further details of clustering and demonstrations of clustering results

Ward linkage was used as the criterion to determine the merges between clusters in agglomerative clustering. This criterion minimises the variance within clusters by merging the pair of clusters with the minimum between-cluster sum of squared differences. Fig. A.6 depicts how many clusters agglomerative clustering selects for neuron 35. The height measures the distance between clusters, so the heights show the distances at which merges occur. The number of clusters C can be inferred by placing a horizontal line at the desired distance or similarity threshold. In the case of this neuron, the dendrogram shows there will be 2 clusters for around 16-28 and more for < 16. At $d_{max} = 15$ we get 3 clusters, if we pick a higher threshold, we would obtain 2 clusters. In step 3, Ward linkage is again used, so the k-means clustering works similarly to agglomerative clustering.

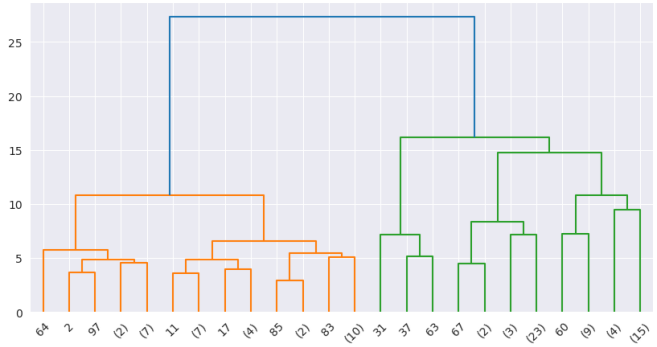


Figure A.6. Neuron 35, agglomerative clustering dendrogram.

Fig. A.7a shows the clustering result for polysemantic neuron 5. This neuron activates highest for images of peacocks and vans. Applying our method yields two disentangled vectors. Fig. A.7b shows the clustering result for monosemantic neuron 13. This neuron activates the most for images with dark backgrounds. Applying our method yields one concept vector. An additional example of a polysemantic neuron is shown in Fig. A.7c. Neuron 27 activates highest for images of toilet paper rolls and the faces of a specific dog breed.

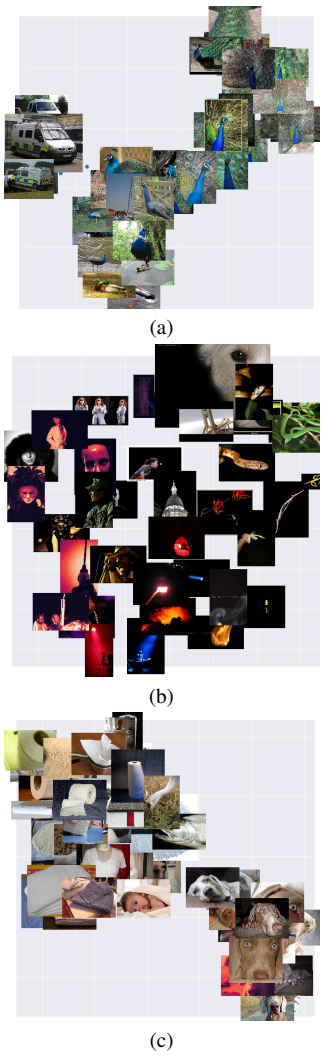
Fig. A.8 depicts the agglomerative clustering dendrogram for neuron 1. Neuron 1 is cut into two clusters for around 14-17 and one for >17. This demonstrates how we can use d_{max} to control how fine-grained the concepts found are.

A.2. Detail on concept vectors

Fig. A.9 shows the two concept vectors found for neuron 35 representing the concepts of apples and sport-type images. It can be seen from the figure that although both categories highly activate neuron 35 that some of the other activation spikes are not common in these two concepts.

A.3. User Evaluation

Fig. A.10 shows examples of questions from our user evaluation test. (a) asks if the concept conveyed by these images is described well by the concept label 'curvy'. One user disagreed with this label, and instead proposed the label 'sinusoidal'.



(a) (b) (c)

Figure A.7. UMAP visualisation of embeddings in latent space corresponding to images kept after k-means clusters and outlier removal for (a) polysemantic neuron 5, (b) monosemantic neuron 13, (c) neuron 1 and (d) neuron 27.

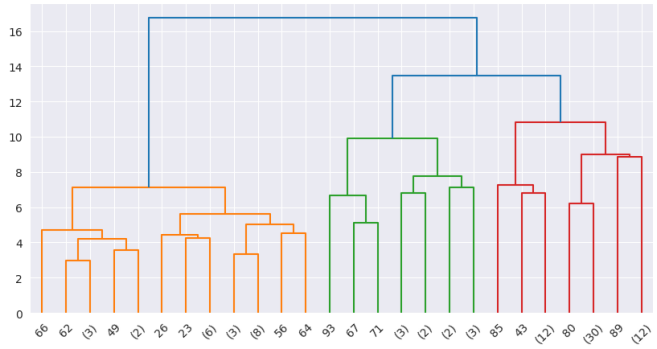
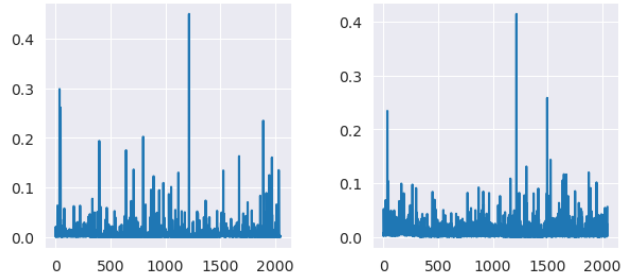
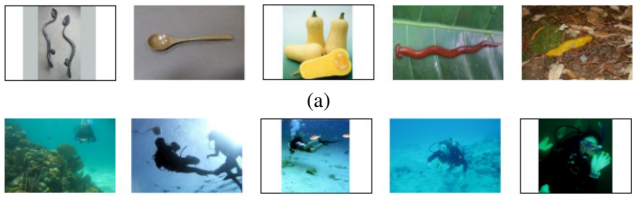


Figure A.8. Neuron 1, agglomerative clustering dendrogram.



(a) Concept vector 1. (b) Concept vector 2.

Figure A.9. Concept vectors found for neuron 35. The x axis represents each dimension in activation space, so each peak is the amount that the concept vector points for the 2048 basis vectors in this particular layer.



(a)

(b)

(c)

Figure A.10. User evaluation questions to evaluate the understandability of the semantic meaning of concepts. For (a) the user was asked if a given label describes the concept well. Participants were asked to label the concept shown in the images in (b). The labels proposed by the users are shown in (c). The font size reflects the frequency of each label suggested by the users.