Supplementary material: Interpretable part-whole hierarchies and conceptual-semantic relationships in neural networks

Figure 1. 2D representation of the latent space for multiple methods trained only on the CIFAR-10 dataset obtained using Principal Component Analysis (PCA) [7]. The PCA provides a deterministic change of base for the data from a multidimensional space into a 2D space. The legend (f) displays the classes, which are divided between super-classes Vehicles and Animals following the WordNet hierarchy [6]. The different methods (a,b,c,d,e) are all able to cluster the samples. However, while (a,b,e) display a latent space where classes are close to each other, the two MLP-based methods (c,d) are able to provide a clearer separation between classes. Both methods show conceptual-semantically close samples on the edge of each superclass, such as airplanes and birds. Inside each superclass, semantically close samples are represented contiguously, such as deers and horses, or cars and trucks. Our method (c) provides better inter-class and intra-class separability. We provide numerical results of the classes overlap in Fig. 2.

A. Content

In the supplementary material we attached:

- Further qualitative results about the latent space organization as the representation of conceptual-semantic relationship in data shown in Fig. 1.
- Additional results in Fig. 2 comparing the correlation between two classes in the latent space.
- A .zip file with the code to run and reproduce our experimental setup and results. The code is written using PyTorch Lightning and it will be included in a GitHub repository upon acceptance.
Figure 2. The overlap percentage $O$ between classes in the latent space is reported for each possible class and each method. For each table, the top-left quadrant represents the overlap percentage between classes belonging to the super-class *animals*, while the bottom-right one for the superclass *vehicles*. The top-right and bottom-left quadrants represent the area where a mistake with a higher hierarchical severity [1] is possible. It would be ideal to have the table with zeros for all the values, but the diagonal. That would represent perfect separation between all the classes. Our method provides the best separation between the two superclasses. It is interesting to notice that the highest intra-superclasses correlation is present between the two classes *bird* and *airplanes* which share features like the wings and the ability to fly.
References


