Disentangling Polysemantic Channels in Convolutional Neural Networks

Supplementary Material

A. Experimental Details

For our experiments, we use the ImageNet dataset [3] and study the popular ResNet-50 architecture [9]. We use pre-trained model weights from PyTorch [19]. Our layer of interest l is the last convolutional layer – earlier layers could have been used as well; however, we chose the last one for simplicity and because we suspect that later layers encoding semantically more meaningful concepts are easier to disentangle with our proposed approach. We consider a channel c as relevant for class t if the condition in Ineq. (2) with $\tau = 0.03$ and $p = 0.75$ is satisfied. We consider a channel affected by polysemanticity if the cosine similarity in Theorem 3.1 is below $\gamma = 0.5$. The values of τ, γ, and p control at what point a channel is considered relevant for a class and at what point it is affected by polysemanticity. Therefore, they are highly application-specific and can be selected more or less conservatively depending on whether the application requires higher recall or higher precision (in this work, τ, γ, and p are handpicked).

More specifically, the hyperparameter γ controls the maximum ARV cosine similarity for which a channel relevant for two classes is considered to be polysemantic. We aim to choose γ to include as many channels as possible while maximizing the chance that the included channels are truly polysemantic and not activating for a shared concept, such as for different dog species. To find an appropriate value for γ, we take all class pairs for which the same channel c is relevant (see Ineq. (2)) and plot the WordNet [4] path similarity for the two class labels over their ARV cosine similarity in Fig. 4. Intuitively, similar classes that share concepts, such as different dog species, have a high WordNet similarity while semantically different classes have a low similarity. Thus, we can use the WordNet similarity as a proxy for semantic similarity, which we use as a proxy for visual similarity [4]. For $\gamma = 0.5$, almost all the relevant class pairs have a very low WordNet similarity, and thus, we continue our analysis with class pairs that have an ARV cosine similarity below that value.

The factor ρ controlling when a channel i in $l-1$ is mainly relevant for the concept of t_2, respectively t_1, in Ineq. (5) is selected automatically. To this end, we take all training images from t_1 and t_2 and measure the activations of the

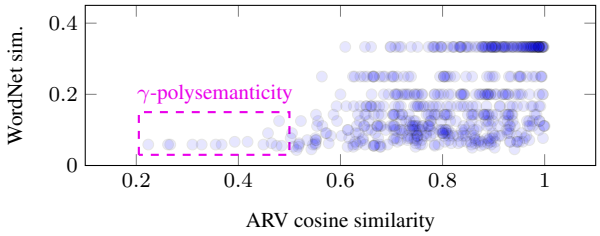

Figure 4. *WordNet similarity of the two classes for which a channel is relevant over their ARV cosine similarity (see Definition 3.1).*

disentangled channels c'_1 and c'_2 for different ρ to maximize

$$\arg\max_{\rho} \frac{1}{|X^{(t_1)}|} \sum_{x \in X^{(t_1)}} \frac{\text{sum}\left(|f_{l'}(x)_{c'_1}|\right)}{\text{sum}\left(|f_{l'}(x)_{c'_2}|\right)} + \frac{1}{|X^{(t_2)}|} \sum_{x \in X^{(t_2)}} \frac{\text{sum}\left(|f_{l'}(x)_{c'_2}|\right)}{\text{sum}\left(|f_{l'}(x)_{c'_1}|\right)}. \tag{6}$$

Qualitative analysis. To visualize the highest activating image patches in Fig. 1 from two classes, we select the 16 highest activating images and extract the highest activating patches. We then choose six patches such that they are equally distributed among the two classes (if there are images from both classes within the 16 patches). This allows us to consider the top 16 patches, without overfilling the plot.

Quantitative analysis. For our quantitative analysis, the relative activation is meaningless for images in which the concept encoded in the channel of interest is absent. Thus, we only include images where the channel of interest has a relative attribution above τ (see Ineq. (2)). With the chosen hyperparameters, this condition applies to at least 75% of the images, representing a sufficiently large number to draw meaningful conclusions.