神经元与概念

> Bau D, Zhou B, Khosla A, et al. Network Dissection: Quantifying Interpretability of Deep Visual Representations[C]// Computer Vision and Pattern Recognition. IEEE, 2017:3319-3327.

量化神经元k与概念c之间的关系：

1. 输入 x，前向传播至k时输出 $A_k(\mathrm{x})$，插值得原输入大小的$S_k(\mathrm{x})$，以阈值T_k得掩码 $M_k(\mathrm{x}) = I(S_k(\mathrm{x}) \geq T_k)$
2. 每个输入x人工标注概念c的掩码 $L_c(\mathrm{x})$
3. 用交并比 $\mathrm{IoU}_{k,c} = \frac{\sum_{\mathrm{x}} M_k(\mathrm{x}) \bigcap L_c(\mathrm{x})}{\sum_{\mathrm{x}} M_k(\mathrm{x}) \bigcup L_c(\mathrm{x})}$ 反映神经元 k 与概念 c 的关系

- What is a disentangled representation, and how can its factors be quantified and detected?
- Do interpretable hidden units reflect a special alignment of feature space, or are interpretations a chimera?
- What conditions in state-of-the-art training lead to representations with greater or lesser entanglement?

emergent interpretability is an axis-aligned property of a representation that can be destroyed by rotation without affecting discriminative power.

1. Identify a broad set of human-labeled visual concepts.
2. Gather hidden variables' response to known concepts.
3. Quantify alignment of hidden variable−concept pairs.