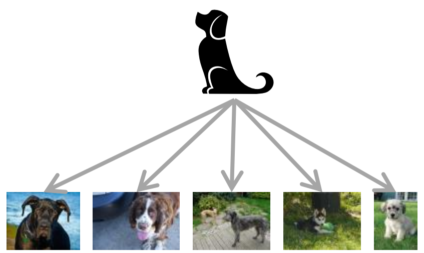
HINT: Hierarchical Neuron Concept Explainer

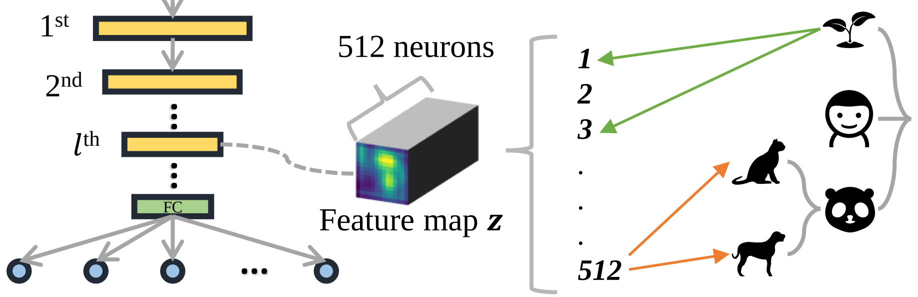
Andong Wang, Wei-Ning Lee, Xiaojuan Qi [Project Video]

I. Introduction

Q: Does the model learn the concept ***dog*** beyond different *breeds of dogs*?



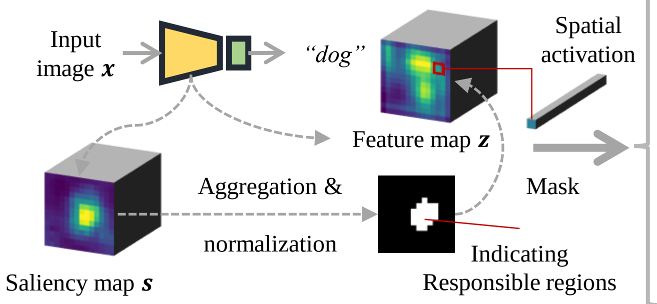
Q: Existing methods ignore the explanation of neurons which are responsible for ***higher-level concepts.***



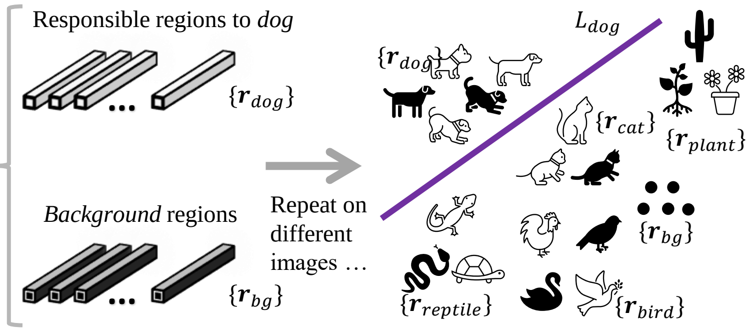
→ ***HINT***: Bidirectional associations between ***neurons*** and ***hierarchical concepts*** (low-cost and scalable)

II. Method

Step 1 *Responsible region* identification.



Step 2 Train *concept classifiers*.

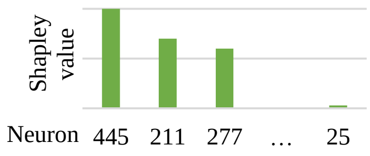


Step 3 Contribution scores of *neurons* to *concepts* (Shapley Values).

Shapley Value ϕ: contribution of neuron d to concept e

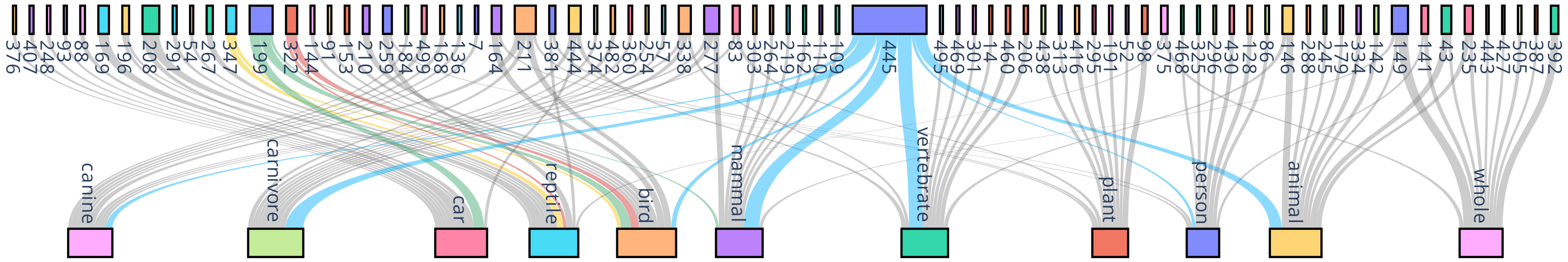
$$\phi = \frac{\sum_r \left|\sum_{i=1}^{M}\left(L_e^{\langle \mathcal{S}\cup d\rangle}(r) - L_e^{\langle \mathcal{S}\rangle}(r)\right)\right|}{M\left|r_{\mathcal{E}} \cup r_{b^*}\right|}$$

Derived from concept classifier L_{dog}



III. Results

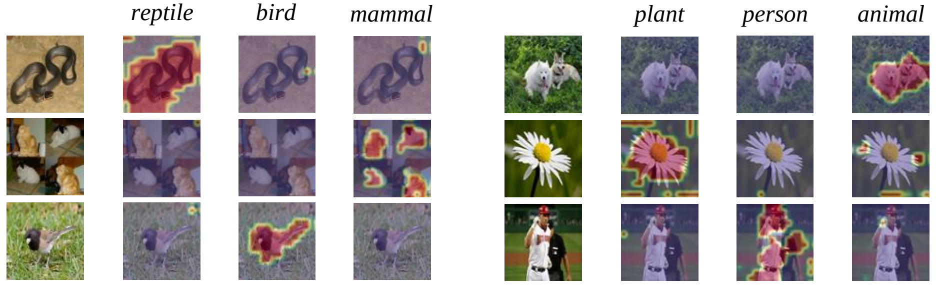
i) Responsible *neurons* to *hierarchical concepts* on layer features.30 of VGG19



ii) Verification of HINT: Weakly Supervised Object Localization (WSOL)

Correct concept classifiers → Correct neuron contributions

Visualization of different concept classifiers applied on ImageNet



Comparison with existing methods on CUB-200-2011 and ImageNet

Table 1. Localization Accuracy on CUB-200-2011.

	VGG16	ResNet50	Inception v3
CAM* [74]	34.4%	42.7%	43.7%
ACoL* [70]	45.9%	-	-
SPG* [71]	-	-	46.6%
ADL* [14]	52.4%	62.3%	53.0%
DANet* [63]	52.5%	-	49.5%
EIL* [37]	57.5%	-	-
PSOL* [66]	66.3%	70.7%	65.5%
GCNet* [33]	63.2%	-	-
RCAM* [6]	59.0%	59.5%	-
FAM* [40]	**69.3**%	**73.7**%	**70.7**%
Ours (10%)	**66.6**%	60.2%	49.0%
Ours (20%)	65.2%	67.1%	55.8%
Ours (40%)	61.3%	77.3%	52.8%
Ours (80%)	64.8%	**80.2**%	**56.2**%

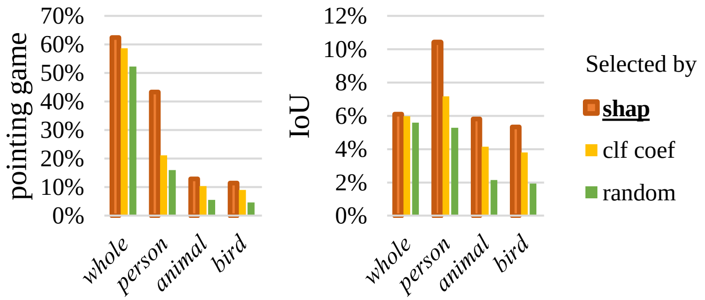
* indicates fine-tuning on CUB-200-2011.

Table 2. Localization Accuracy on ImageNet.

	VGG16	ResNet50	Inception v3
CAM [74]	42.8%	-	-
ACoL [70]	45.8%	-	-
SPG [71]	-	-	48.6%
ADL [14]	44.9%	48.5%	48.7%
DANet [63]	-	-	48.7%
EIL [37]	46.8%	-	-
PSOL [66]	50.9%	54.0%	54.8%
GCNet [33]	-	-	49.1%
RCAM [6]	44.6%	49.4%	-
FAM [40]	**52.0**%	**54.5**%	**55.2**%
Ours (10%)	64.7%	59.7%	53.1%
Ours (20%)	**66.1**%	66.6%	54.1%
Ours (40%)	64.4%	69.4%	54.3%
Ours (80%)	62.6%	**70.7**%	**58.7**%

iii) Effectiveness of Shapley Values

Train concept classifiers with 20 neurons



IV. Conclusions

- ***HINT*** presents the first attempt to associate *neurons* with *hierarchical concepts*, which enables us to systematically and quantitatively study whether and how the neurons learn the high-level hierarchical relationships of concepts implicitly.
- ***HINT*** achieves remarkable performance in a variety of applications (see the main paper).

https://github.com/AntonotnaWang/HINT