Concept Generalization in Visual Representation Learning

NAVER LABS Europe
Inria
Mert Bulent Sariyildiz Yannis Kalantidis Diane Larlus Karteek Alahari

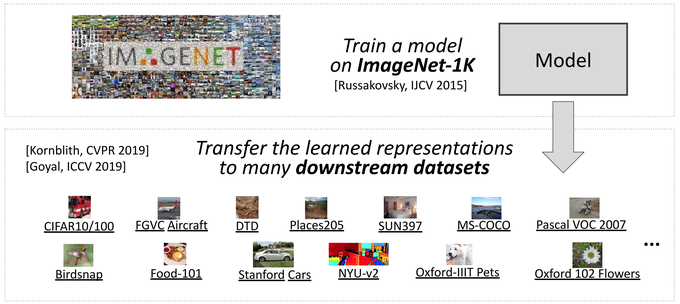
https://europe.naverlabs.com/cog-benchmark

2021 ICCV OCTOBER 11-17 VIRTUAL

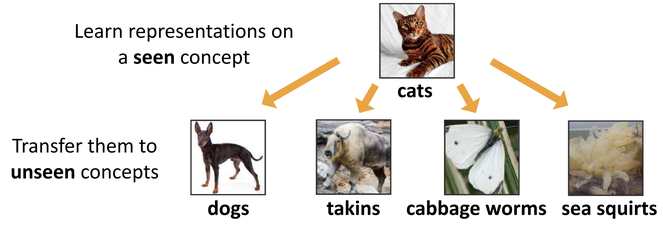
Highlights / Take-home messages

- ➔ **We propose the ImageNet-CoG benchmark**
  - ◆ Enables measuring concept generalization in a principled way
  - ◆ **Seen** concepts ⇒ ImageNet-1K concepts
  - ◆ **Unseen** concepts ⇒ Sampled from the full ImageNet-21K dataset
    - ● **5 Levels** ⇒ Increasingly challenging transfer datasets
- ➔ **To be used out-of-the-box** for ImageNet-1K pretrained models
- ➔ **31 models evaluated** on ImageNet-CoG
  - ◆ Interesting insights on popular state-of-the-art methods

Learning general-purpose visual representations



Concept generalization



- ➔ No systematic approach for evaluating concept generalization
- ➔ Unknown semantic similarity between ImageNet-1K and other datasets

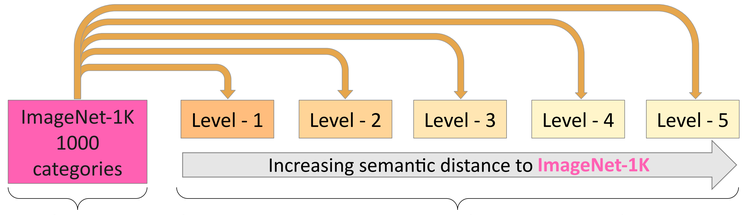
We tackle the following questions:

- How can we evaluate concept generalization reliably?
- Which methods are the best for concept generalization?

Proposed ImageNet-CoG Benchmark

A benchmark tailored for concept generalization, built on the full ImageNet-21K [Deng, CVPR 2009]

➔ Transfer learning scenarios from ImageNet-1K to ImageNet-CoG levels (sampled from the full ImageNet)

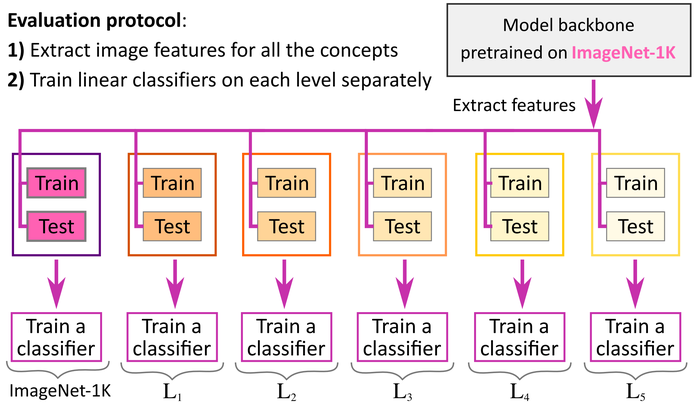


- ➔ Seen and unseen concepts are in the same concept ontology (WordNet ontology [Miller, ACM 1995])
- ➔ Semantic similarity between concepts is defined by linguists

Evaluation protocol:

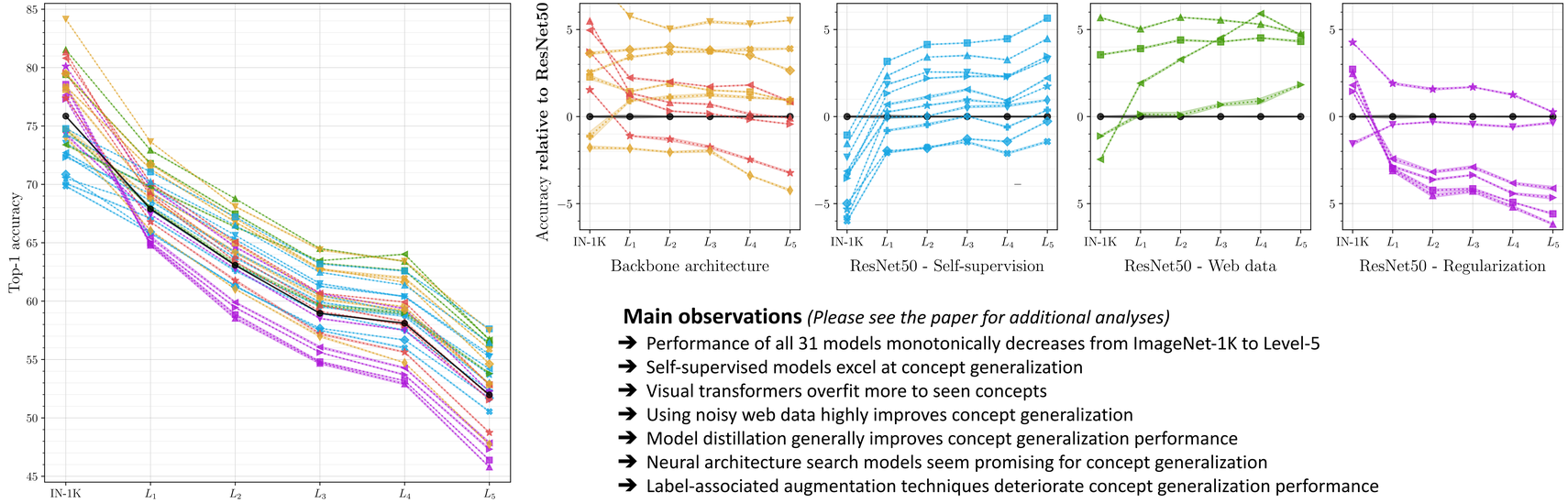
1) Extract image features for all the concepts

2) Train linear classifiers on each level separately



Main results of evaluating 31 representation learning methods on the ImageNet-CoG Benchmark

Baseline	Backbone architecture		Self-supervision		Web data	Regularization	
● ResNet50 (23.5M)	▲ *a*-T2T-ViT-t-14 (21.1M)	★ *a*-Inception-v3 (25.1M)	■ *s*-DINO	★ *s*-SimCLR-v2	■ *d*-Semi-Sup	■ *r*-ReLabel	▼ *r*-Adv-Robust
	▶ *a*-DeiT-S (21.7M)	✚ *a*-EfficientNet-B1 (6.5M)	▲ *s*-SwAV	✚ *s*-MoCo-v2	▲ *d*-Semi-Weakly-Sup	▲ *r*-CutMix	★ *r*-MEAL-v2
	◀ *a*-DeiT-S-distilled (21.7M)	✖ *a*-EfficientNet-B4 (17.5M)	▶ *s*-BarlowTwins	✖ *s*-MoCHi	▶ *d*-MoPro	▶ *r*-MixUp	
	▼ *a*-DeiT-B-distilled (86.1M)	◆ *a*-NAT-M4 (7.6M)	◀ *s*-OBoW	◆ *s*-CompReSS	◀ *d*-CLIP	◀ *r*-Manifold-MixUp	
	■ *a*-ResNet152 (58.1M)	♦ *a*-VGG19 (139.6M)	▼ *s*-BYOL	♦ *s*-InfoMin			



Main observations *(Please see the paper for additional analyses)*

- ➔ Performance of all 31 models monotonically decreases from ImageNet-1K to Level-5
- ➔ Self-supervised models excel at concept generalization
- ➔ Visual transformers overfit more to seen concepts
- ➔ Using noisy web data highly improves concept generalization
- ➔ Model distillation generally improves concept generalization performance
- ➔ Neural architecture search models seem promising for concept generalization
- ➔ Label-associated augmentation techniques deteriorate concept generalization performance