The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition

Jonathan Krause[1], Benjamin Sapp[2], Andrew Howard[2], Howard Zhou[2], Alexander Toshev[2], Tom Duerig[2], James Philbin[2], Li Fei-Fei[1]

[1]Stanford University, [2]Google

Research at Google

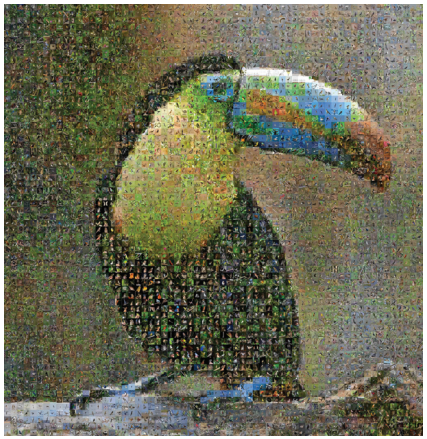
Problem

-Fine-grained recognition works well with labels

-But fine-grained labels are **expensive**

-There are too many fine-grained categories in the world to annotate by hand: 14k birds, 278k butterflies and moths, 941k insects

-How can we scale up fine-grained recognition?



4,224 (+1) categories recognized in this work

Contributions

-Demonstrate feasibility of training models of fine-grained with noisy data from the web and simple, generic, models of recognition.

-Greatly improved recognition performance on four fine-grained datasets without using ground truth training data.

-Scale fine-grained recognition to over 10,000 species of birds and 14,000 species of butterflies and moths.

Data

Categories

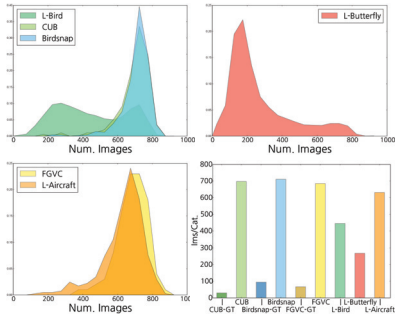
Birds: 10,982 species

Butterflies: 14,553 species (+moths)

Aircraft: 409 varieties

Dogs: 515 breeds

-Images from Google Image Search

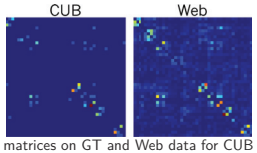


Noise

Cross-domain noise: portion of images that are not of any fine-grained category in a given domain. Measure by hand.

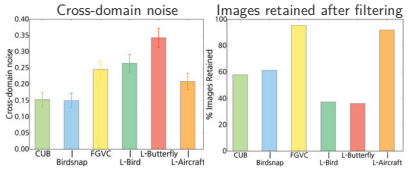
Cross-category noise: portion of images that have the wrong fine-grained label. Hard to estimate.

Filtering: Proposed technique to reduce noise. Simply remove images with multiple web labels!



Confusion matrices on GT and Web data for CUB categories

Examples of cross-domain noise

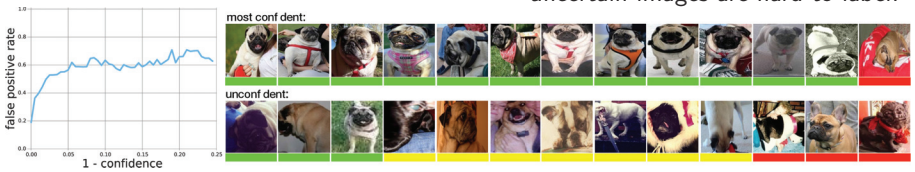


Images removed after filtering and their web labels

Active Learning

Alternative approach for collecting large quantities of fine-grained data.

We use *confidence-based sampling*: Label the most confident images. Can still fix false positives, and uncertain images are hard to label!



Experiments

-Use Inception-v3 CNN classifier

-Extensive dedup with ground truth test datasets via [2]

-YFCC100M data for active learning

Training Data	Acc.	Dataset
CUB-GT	84.4	CUB
Web (raw)	87.7	CUB
Web (filtered)	89.0	CUB
L-Bird	91.9	CUB
L-Bird(MC)	92.3	CUB
L-Bird+CUB-GT	92.2	CUB
L-Bird+CUB-GT(MC)	92.8	CUB
Birdsnap-GT	78.2	Birdsnap
Web (raw)	76.1	Birdsnap
Web (filtered)	78.2	Birdsnap
L-Bird	82.8	Birdsnap
L-Bird(MC)	85.4	Birdsnap
L-Bird+Birdsnap-GT	83.9	Birdsnap
L-Bird+Birdsnap-GT(MC)	85.4	Birdsnap

Training Data	Acc.	Dataset
FGVC-GT	88.1	FGVC
Web (raw)	90.7	FGVC
Web (filtered)	91.1	FGVC
L-Aircraft	90.9	FGVC
L-Aircraft(MC)	93.4	FGVC
L-Aircraft+FGVC-GT	94.5	FGVC
L-Aircraft+FGVC-GT(MC)	95.9	FGVC
Stanford-GT	80.6	Stanford Dogs
Web (raw)	78.5	Stanford Dogs
Web (filtered)	78.4	Stanford Dogs
L-Dog	78.4	Stanford Dogs
L-Dog(MC)	80.8	Stanford Dogs
L-Dog+Stanford-GT	84.0	Stanford Dogs
L-Dog+Stanford-GT(MC)	85.9	Stanford Dogs

Prior work on GT datasets:

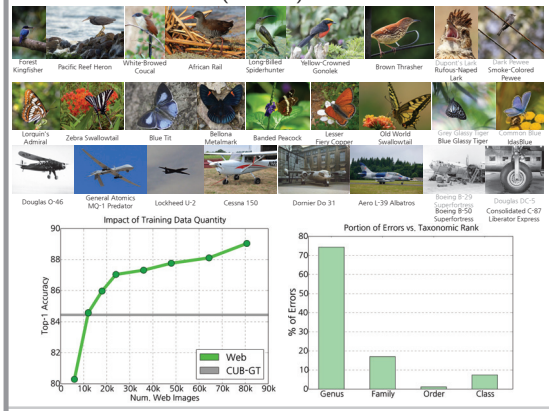
CUB: 84.6% (Xu et al. ICCV'15)
Birdsnap: 66.6% (Berg et al. CVPR'14)
FGVC: 84.1% (Lin et al. ICCV'15)
Stanford Dogs: 76.8% (Sermanet et al. ICLR'2015)

Training Procedure	Acc.
Stanford-GT (scratch)	58.4
A.L., one round (scratch)	65.8
A.L., two rounds (scratch)	74.0
Stanford-GT (fine-tune)	80.6
A.L., one round (ft)	81.6
A.L., one round (ft, subsample)	78.8
A.L., two rounds (ft)	82.1
Web (filtered)	78.4
Web (filtered) + Stanford-GT	82.6

Very Large-Scale Fine-Grained Recognition

-Test on Flickr images w/exact category name matches, deduped with other web images.

-Accuracy: Birds (73.1%), Butterflies (65.9%), Aircraft (72.7%)



References

[1] Szegedy et al. Rethinking the Inception Architecture for Computer Vision. CVPR 2016
[2] Wang et al. Learning Fine-Grained Image Similarity with Deep Ranking. CVPR 2014
[3] Wah et al. The Caltech-UCSD Birds-200-2011 Dataset. Tech. Report 2011
[4] Berg et al. Birdsnap: Large-Scale Fine-Grained Visual Classification of Birds. CVPR 2014
[5] Maji et al. Fine-Grained Visual Classification of Aircraft. Tech. Report 2013
[6] Khosla et al. Novel Dataset for Fine-Grained Classification. FGVC 2011