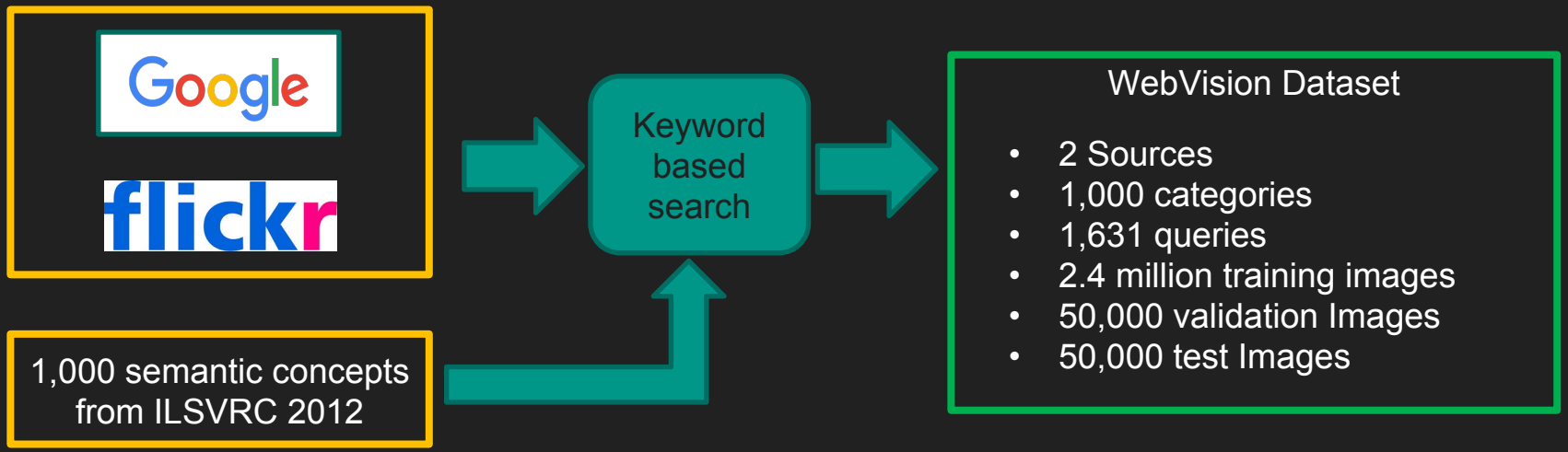
Database Overview

WebVision Database: Visual Learning and Understanding from Web Data,
Wen Li, Limin Wang, Erikur Agustsson, and Luc Van Gool, arXiv 1708.02862, 2017.

webvision

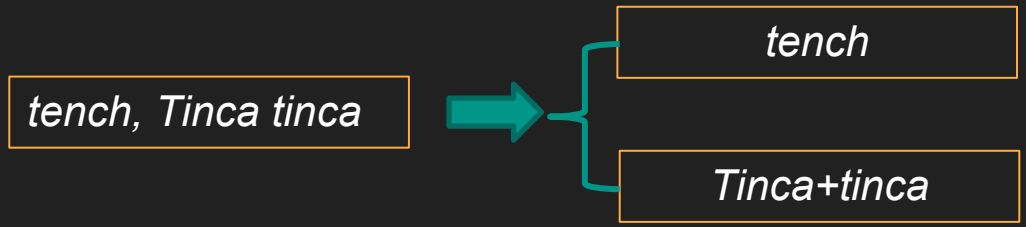
Dataset Construction



Synset to Queries

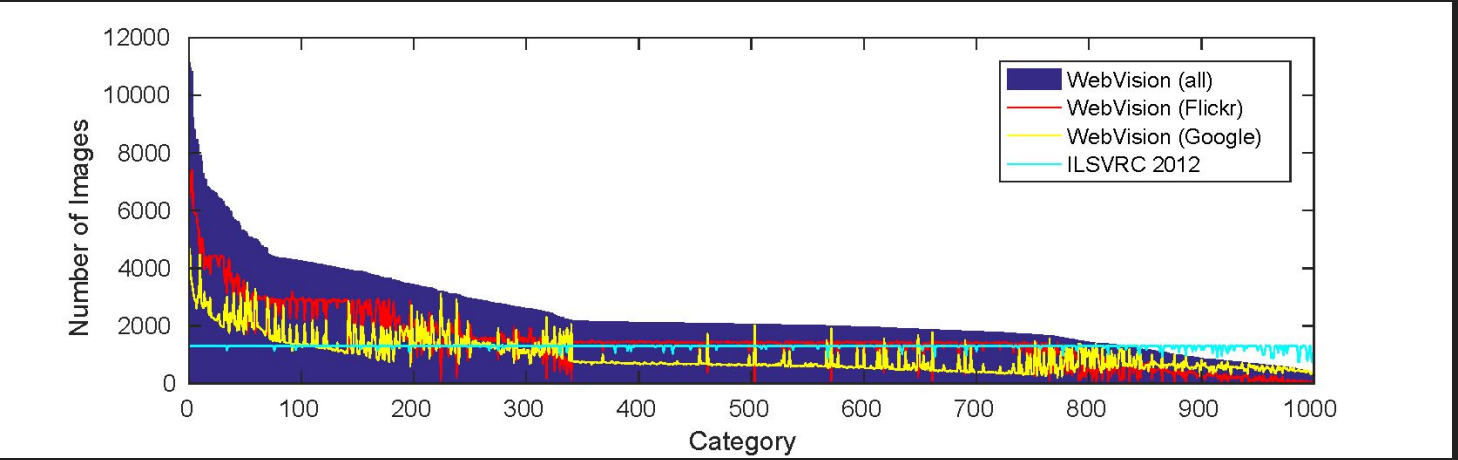
We split each synset into multiple words, and use each word as one query

Overlapped and ambiguous queries are manually removed.

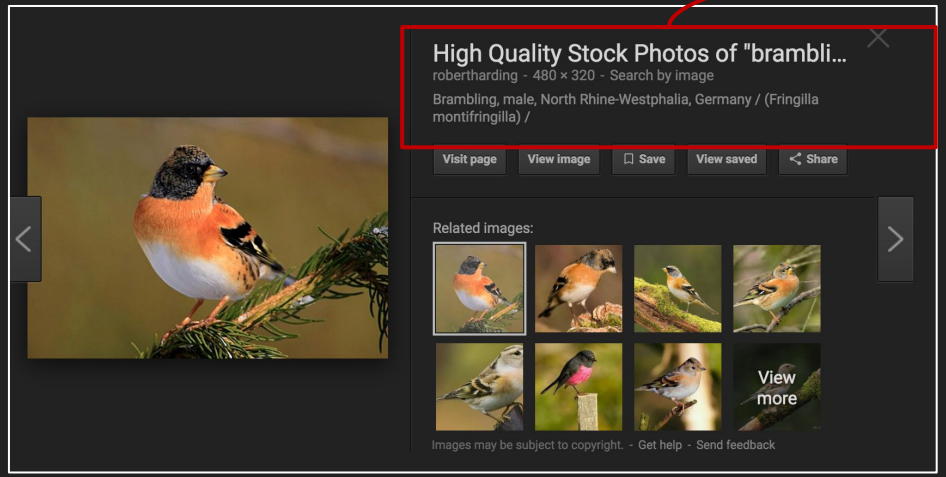


Class distribution

#images per class varies, subject to #queries per class and the availability of images, but generally more than ImageNet

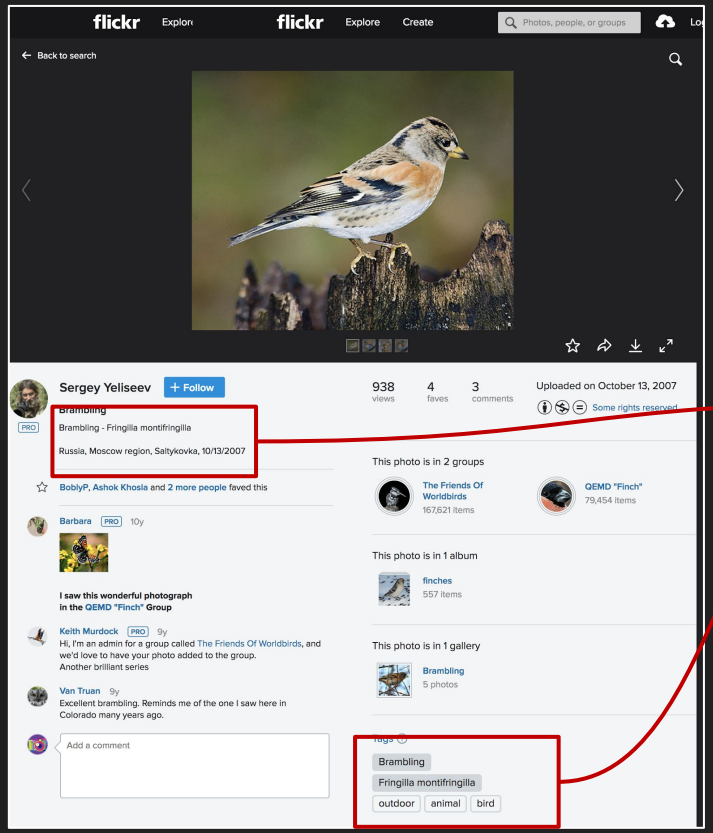


Meta Information - Google Images



- Title: ``*High Quality Stock Photos of brambling*";
- Description: ``*Brambling, male, North Rhine-Westphalia, Germany (Fringilla montifringilla)*";

Meta Information - Flickr Images



- Title: ``*Brambling*";
- Description:``*Brambling - Fringilla montifringilla Russia, Moscow region, Saltykovka, 10/13/2007*";
- Tags: "*Brambling*", "*Fringilla montifringilla*";

Noise

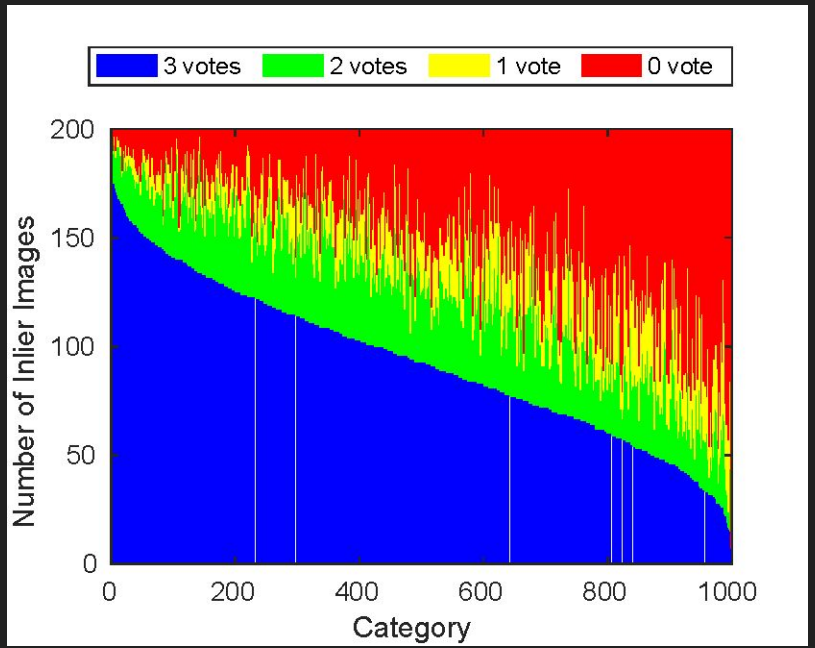
Ask users if the image is correctly labeled or not.

Each Image is annotated by three users.

> 2 votes, 66%

“867-Tractor”, 199/200 inliers

“627-lighter...”, 24/200 inliers



Baseline Results

We use the AlexNet model.

	ILSVRC 2012 Val	**WebVision Val**
ILSVRC 2012	79.77 (56.79)	74.64 (52.58)
WebVision	70.36 (47.55)	77.90 (57.03)

Quantity vs. Quality

#images matters, reducing training images leads to worse results.

The trends show that increasing training images could gain further improvement

Compare when using similar number of training images

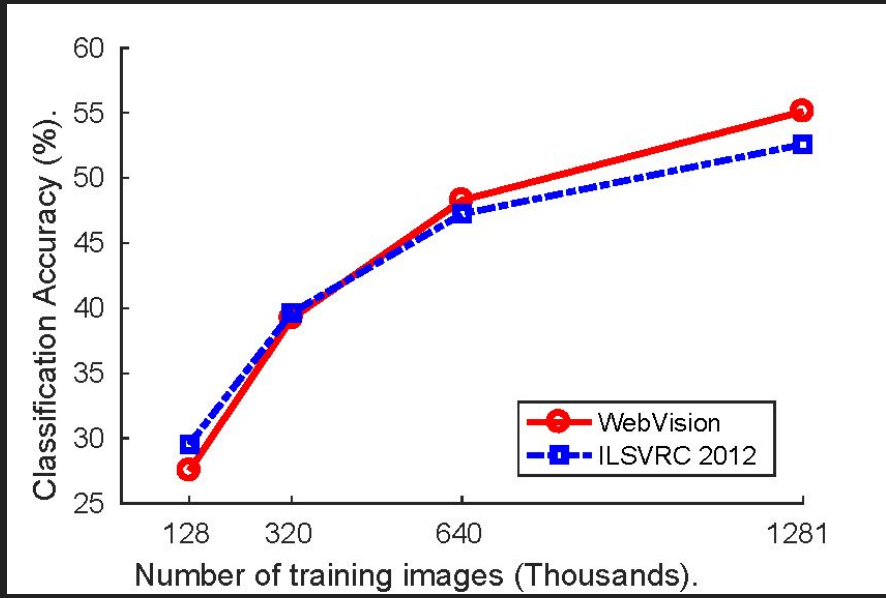
- 50% of WebVision is still similar to 100% of ImageNet
- 10% of WebVision is far worse than 25% of ImageNet

Quantity vs. Quality

Sampling exact the same #training images.

For large number of training images, webvision is better than ImageNet (domain advantage)

For small number of training images, webvision becomes worse, noise images do matter.



Transfer to New Tasks - Image Classification

Use AlexNet models learnt from WebVision and ImageNet to extract features. SVM is used to learn classifiers.

	Caltech 256	**PASCAL VOC 2007**
ILSVRC 2012	70.44	75.65
WebVision	70.43	77.78
Combined	73.61	78.46

Transfer to New Tasks - Object Detection

Finetune AlexNet models learnt from WebVision and ImageNet for the new task.

Model	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
ILSVRC 2012	64.0	67.3	53.5	39.6	32.4	66.6	71.6	66.4	36.6	61.1	58.4	62.2	75.1	70.4	65.0	33.1	56.5	48.0	69.3	64.4	**58.1**
WebVision	64.6	70.6	50.8	41.8	28.6	66.8	71.4	69.4	34.6	63.2	61.8	62.1	74.4	69.7	65.1	32.8	53.2	52.2	70.8	59.5	**58.2**

Summary

- A new web image dataset calibrated with the ILSVRC 2012 dataset, no human annotation for images
- Models trained on the WebVision dataset achieved similar top-5 accuracy
- Domain adaptation issue is observed between two datasets
- The label noise issue could be reduced to some extent with a large number of training images
- When being transferred to new tasks, the model learnt from WebVision achieved comparable results as the ILSVRC 2012 model