Database Overview

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

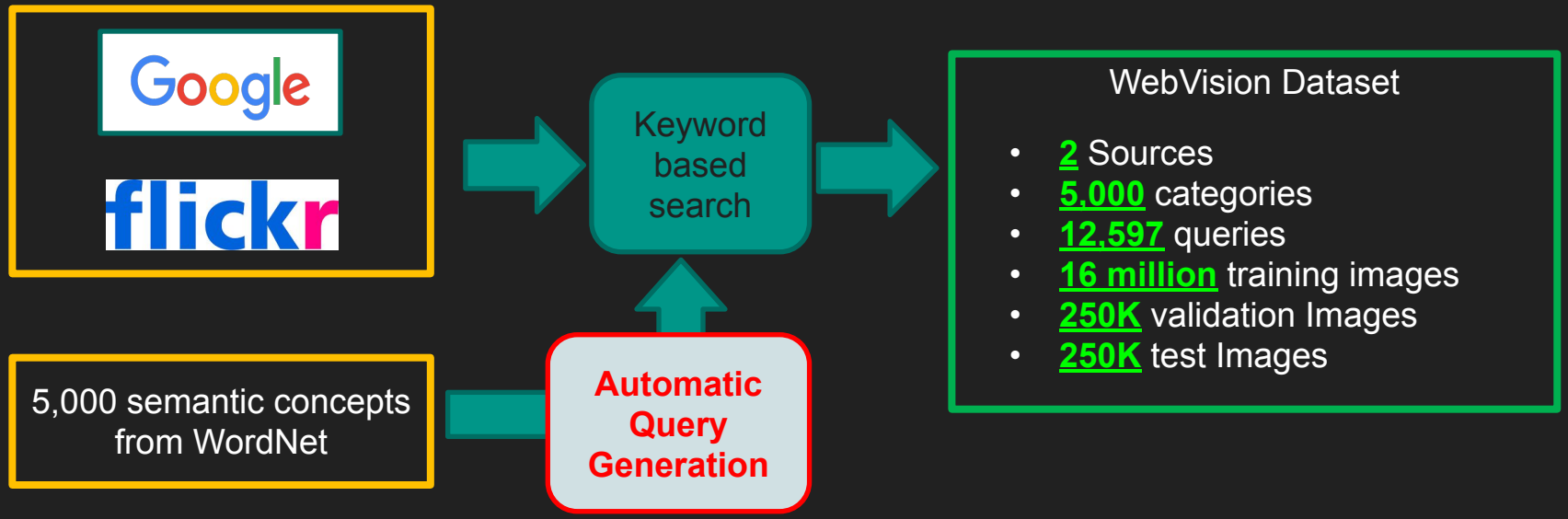
webvision

WebVision2.0 dataset

- **5,000** categories
- From Flickr & Google
- **16M** images
- **290K** validation images
- **290K** test images

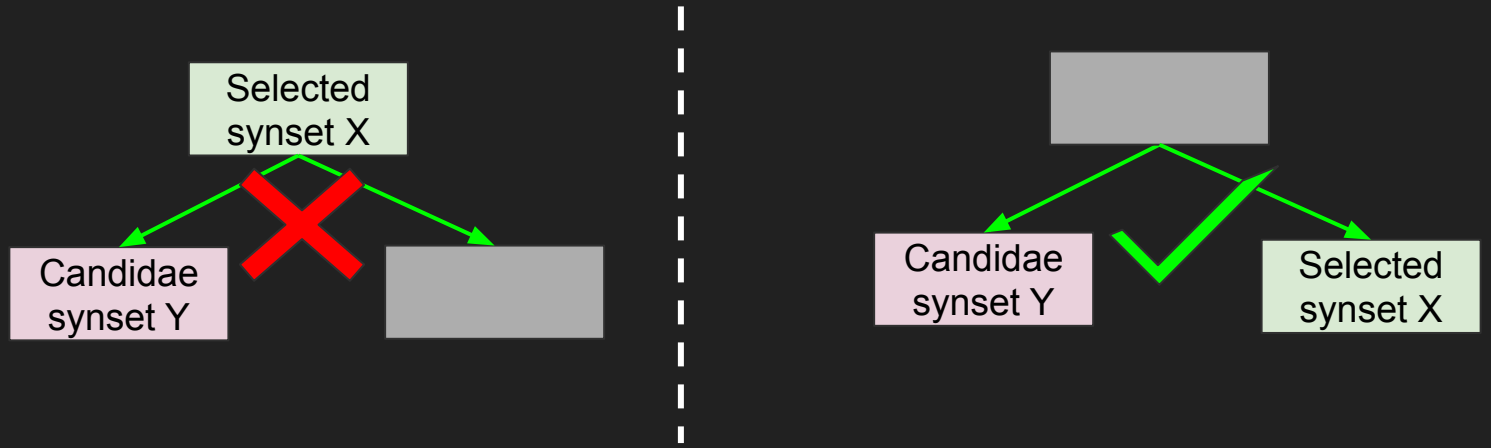
Dataset Construction

Automatic query generation instead of manual way



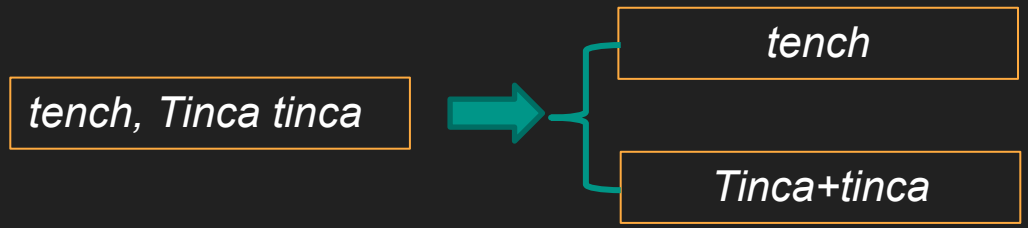
5000 Synsets

- Synsets from ILSVRC2012 dataset are the first 1,000 synsets
- The other 4,000 synsets are selected as follows
  - Sort the remaining synsets in WordNet in descending order according to popularity (the number of images in ImageNet)
  - A synset is valid if and only if it does not cause semantic overlap, i.e., there is not selected synset that is the ancestor node or child node of this synset in WordNet.



Synset to Queries

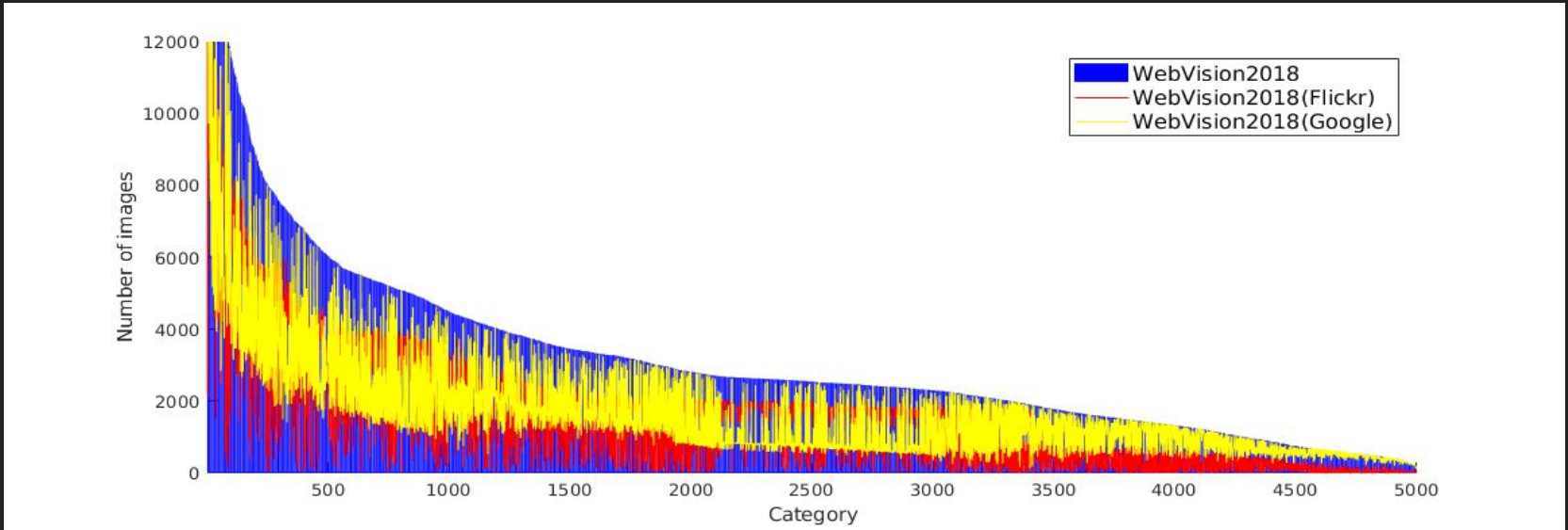
- Synsets are processed in order
- Each synset is splitted into multiple words, and each word is a query
- If a query is overlapped with existing queries, it will be discarded
- If no query is valid for a synset, we combine each word with each word in its parental node to get extended queries.
- If none of those extended query is valid, we discard this synset.
- In total, we get 12,597 queries for 5,000 synsets



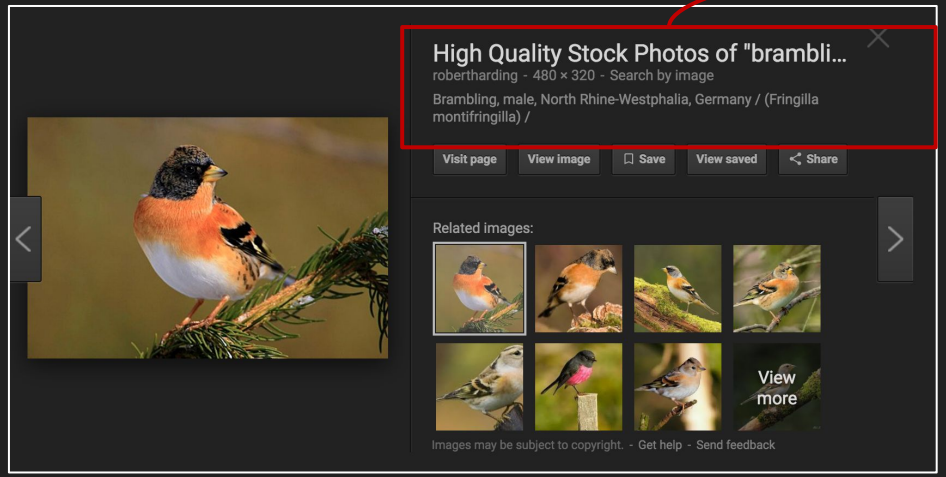
Class distribution

Highly imbalanced

#images/class varies, subject to #queries/class and the availability of images

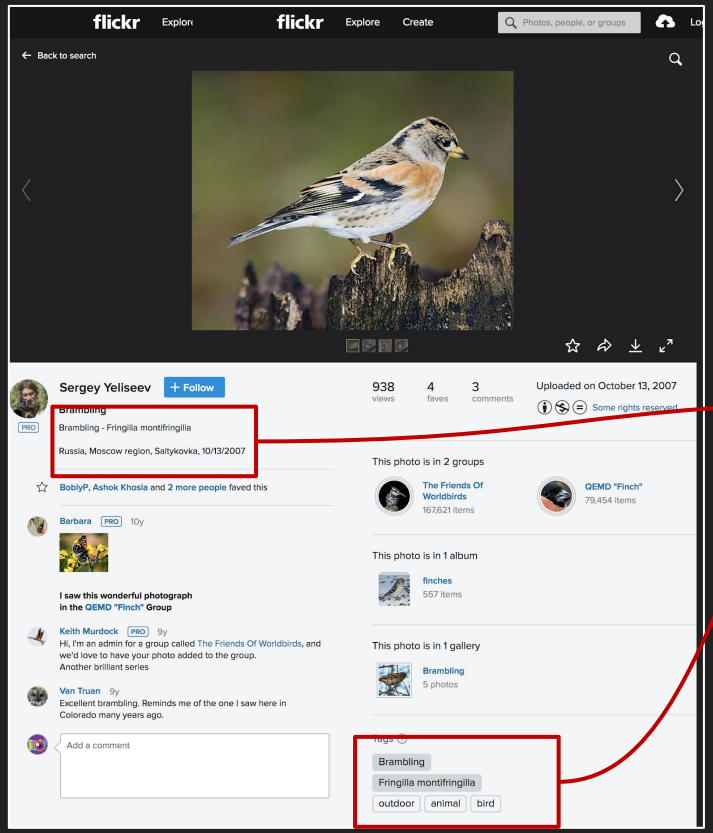


Meta Information - Google Images



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

Meta Information - Flickr Images



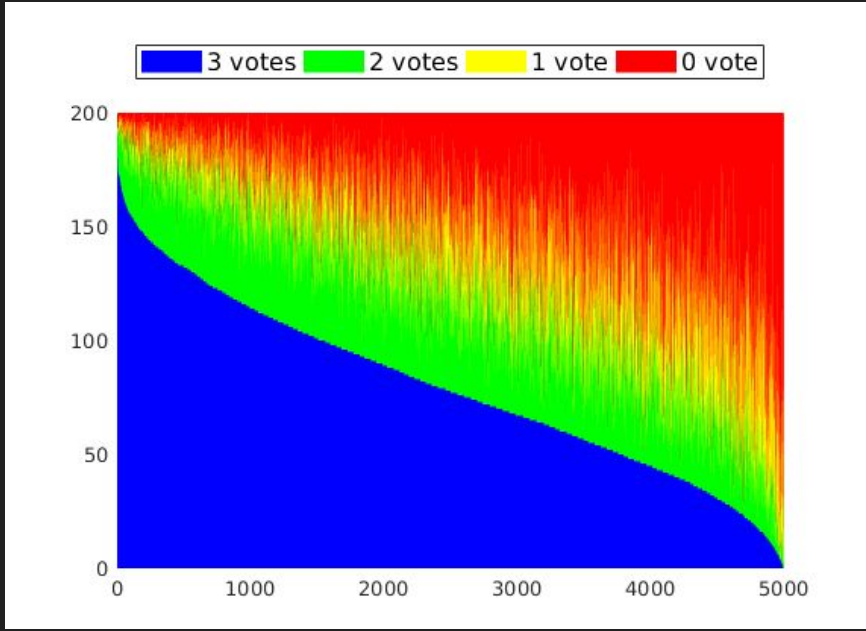
- Title: ``*Brambling*";
- Description:``*Brambling - Fringilla montifringilla Russia, Moscow region, Saltykovka, 10/13/2007*";
- Tags: "*Brambling*", "*Fringilla montifringilla*";

Noise

Ask users if the image is correctly labeled or not.

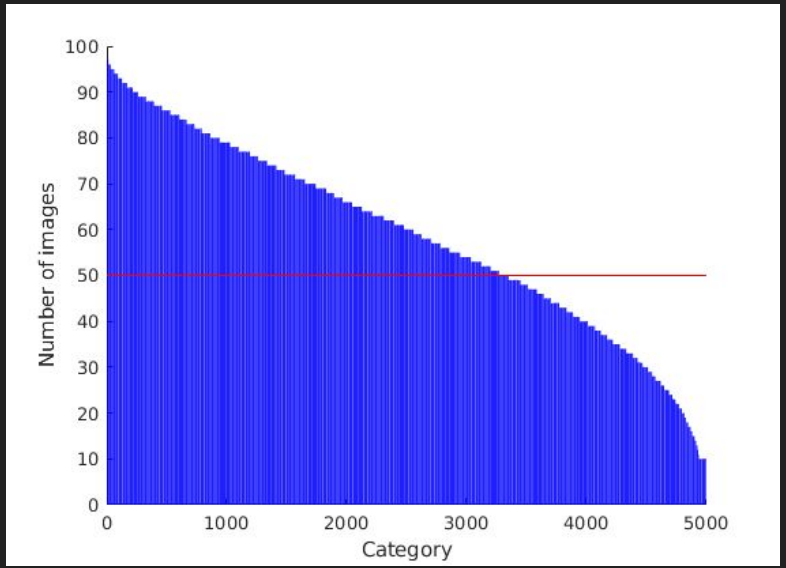
Each Image is annotated by three users.

About 59% images are inliers (with at least 2 votes).



Validation and Test Sets

- Inlier images are highly imbalanced among different classes.
- We preserve this natural imbalance in web images.
- Evenly splitting inlier images into two sets, leading to our validation and test sets.



Evaluation Metric

Due to the imbalance in number of images per class in the val/test set, we use the mean of per class top-5 accuracy as the evaluation metric,

$$ACC = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c} acc(\mathbf{p}_i, y_i)$$

Summary

- A large scale web image dataset with 16M images from 5,000 categories.
- Automatic query generation from WordNet synset
- Preserve the nature of images in the wild:
  - Noisy labels,
  - imbalanced training data
  - imbalanced validation/test data
- Meta information is available

Challenge Overview

Challenge Task

WebVision Image Classification Task

- Learn models on the WebVision train set and evaluate on the val and test set

Challenge Platform

WebVision Challenge 2019

Organized by 07wanglimin - Current server time: June 14, 2019, 2:46 a.m. UTC

First phase	End
Development	Competition Ends
March 1, 2019, midnight UTC	June 8, 2019, 6:59 a.m. UTC

Learn the Details | Phases | Participate | Results | Forums

- Overview
- Evaluation
- Terms and Conditions
- Get Starting

Challenge

The goal of this challenge is to advance the area of learning knowledge and representation from web data. The web data not only contains huge numbers of visual images, but also rich meta information concerning these visual data, which could be exploited to learn good representations and models. In 2019, we organize one track for this challenge: WebVision Image Classification Task.

Challenge Schedule

Development

Start: March 1, 2019, midnight

Description: The Development Leaderboard is based on a fixed random subset of 50% of the test images. To submit, upload a .zip file containing a predictions.txt file with the prediction in the format used in the dev kit. An example submission file can be found at: http://www.vision.ee.ethz.ch/webvision/files/example_submission.zip

Testing

Start: June 1, 2019, midnight

Description: To submit, upload a .zip file containing a predictions1.txt, ..., predictions5.txt file with the prediction in the format used in the dev kit. The file with the best top-5 accuracy will be used to determine the winner. Please also include a readme.txt file with a description for your entry. An example submission file can be found at: http://www.vision.ee.ethz.ch/webvision/files/example_submission_testphase.zip

Competition Ends

June 8, 2019, 6:59 a.m.

Submission Policies

- Each participant may have maximum 10 submissions during development phase.
- Each team may have 1 submissions (containing 5 predictions) during test phase.
- Learn vision models from noisy data (WebVision dataset).
- No extra data is allowed to use.

VQA Web

Frequently Asked Questions

- **Can I use the ImageNet images or the ImageNet pretrained models?**
  No. The main target of WebVision challenge is to push the envolope of learning visual representation without human annotations. So human annotated data is strictly prohibited to be used (Text data will be an exception). Therefore, ImageNet images or ImageNet pretrained models are not allowed to be used in any form.

- **Can I use external images without human annotations?**
  No. For fairness, we restrict the challenge to use only WebVision training images. You are not allowed to use other web image datasets like YFCC100. You are not allowed to crawl web images by yourself, too.

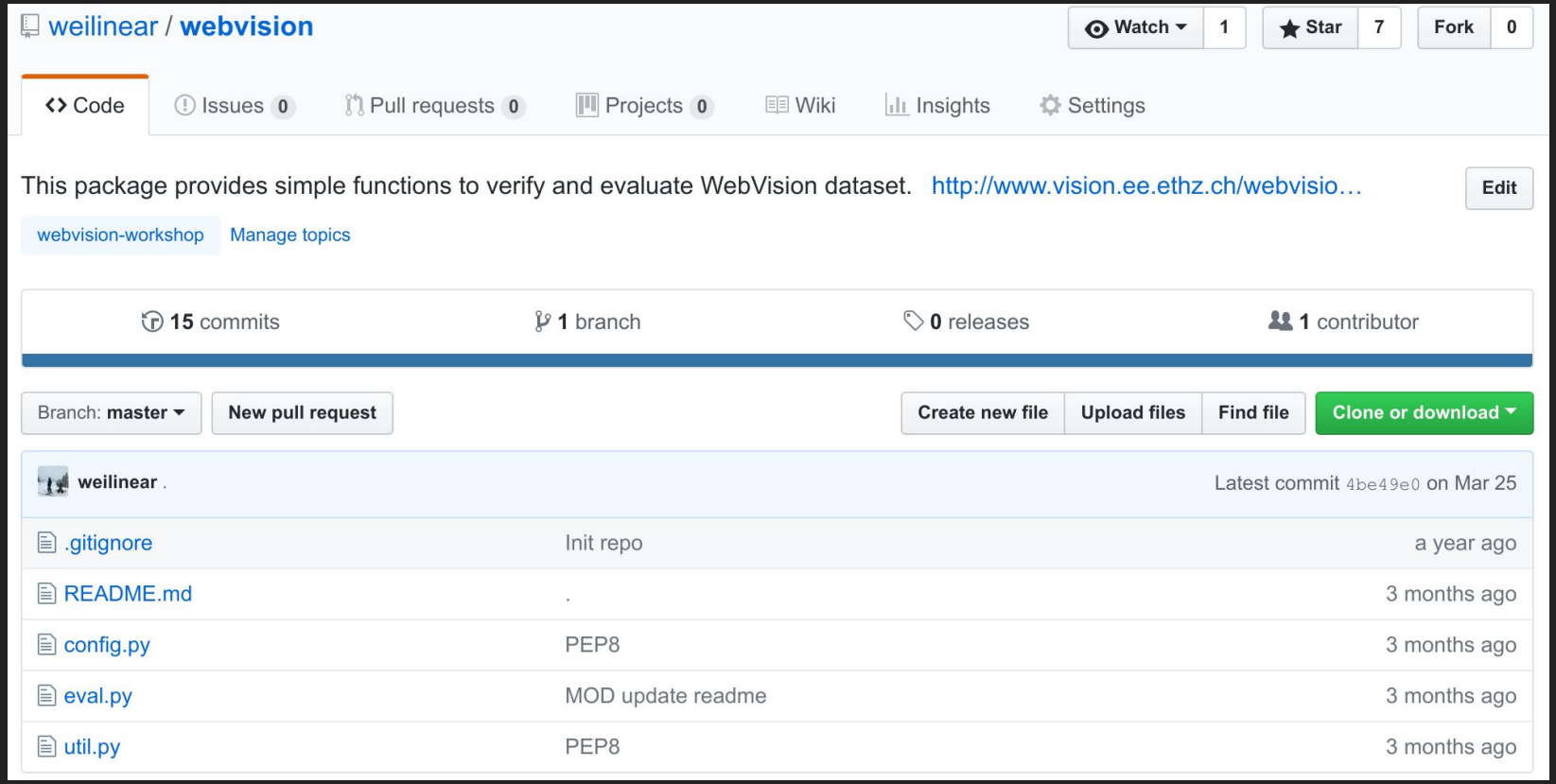
- **Can I use the text data (tags, description, caption) in the WebVision dataset?**
  Yes, and we encourage you to do so. It has shown in the literature that such textual information could provide useful supervision for training models.

- **Can I use external text data, or models pretrained with external text data, with or without human annotation?**
  Yes, and we also encourage you to do so. This does not conflict with our target of learning visual representation without human annotationos. Therefore, WordNet, Knowledge Graph, etc. can be used. Models trained using external text data are also allowed, such as Word2Vec, BERT models, and so on. **Note that the text data or models should be pulicly available. You should explicitly state in your final submission that what text datasets/models are used.**

- **Can I crawl text data according to WebVision concepts by myself, and use it as training data?**
  Yes. There is no restriction on non-visual data except the data should be publicly available. So people could reproduce the results. If you crawl text data by yourself, please clearly state it in your submission, and make it available to public before the final submission deadline. An URL should be provided in the method description part of your submission.

If you have other questions, please drop an email to webvisionworkshop AT gmail.com

Provided Tools

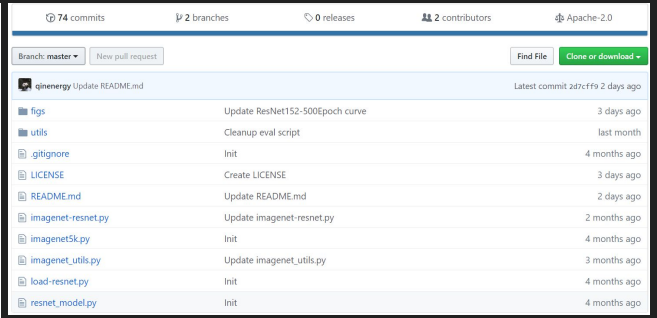


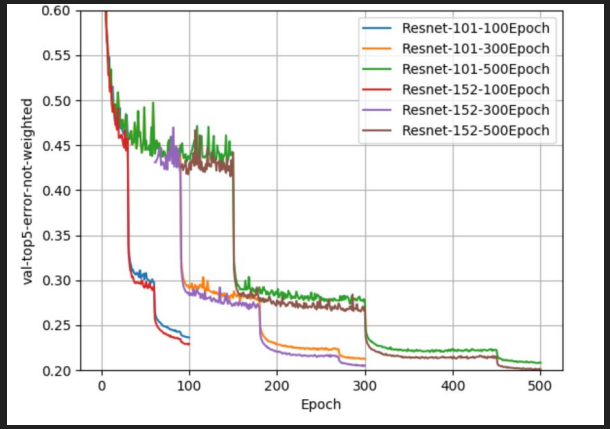
Baseline

Pretrained models

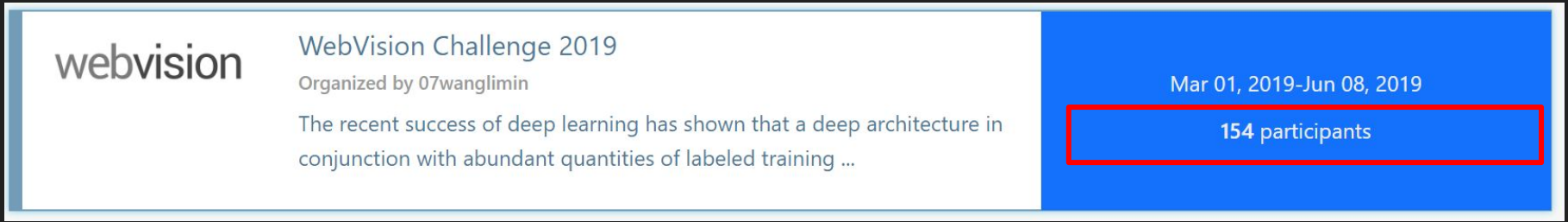
We offer several pretrained models. **Due to the class imbalance in WebVision, we duplicated the file items in train.txt such that different classes have equal number of training samples. You might want to add similar strategies in imagenet5k.py or modify your own train.txt. Check utils/upsample.py for an example.**

Model	Top1-Val-Error	Top5-Val-Error	Download
ResNet-50 (101 Epoch)	54.28%	30.69%	link
ResNet-50 (205 Epoch)	52.10%	28.51%	link
ResNet-101 (100 Epoch)	52.21%	28.62%	link
ResNet-101 (200 Epoch)	50.12%	26.78%	link
ResNet-101 (300 Epoch)	48.97%	25.74%	link
ResNet-101 (500 Epoch)	48.38%	25.21%	link
ResNeXt-101 (100 Epoch)	50.62%	27.11%	link
ResNet-152 (100 Epoch)	51.23%	27.80%	link
ResNet-152 (200 Epoch)	48.98%	25.75%	link
ResNet-152 (300 Epoch)	48.05%	24.88%	link
ResNet-152 (500 Epoch)	47.31%	24.31%	link
ResNet-152-SE (100 Epoch)	51.61%	28.02%	link





Number of participants



We have 5 teams to submit valid results to image classification track.

Challenge Results

Rank	Team name	Best Top-5 Accuracy (%)	Entry-1	Entry-2	Entry-3	Entry-4	Entry-5
1	**Alibaba-Vision**	**82.54**	82.37 / 60.03	82.50 / 60.13	82.51 / 60.22	82.54 / 60.24	82.54 / 60.24
2	**BigVideo**	**82.05**	82.01 / 59.77	82.02 / 59.73	82.01 / 59.73	81.94 / 59.66	82.05 / 59.80
3	**huaweicloud**	**81.15**	80.46 / 57.74	81.07 / 58.54	81.15 / 58.60	81.11 / 58.60	81.15 / 58.63
4	Y_Y	80.69	80.69 / 57.88	80.61 / 57.87	80.45 / 57.09	79.75 / 56.57	80.50 / 57.49
5	PCI	77.92	76.64 / 54.85	77.17 / 55.57	77.20 / 55.51	77.92 / 55.88	75.18 / 52.82

Team: Alibaba-Vision

Modalities: Image, Query ID, Text

The main idea of our method is to learn with side information provided by search engine, WordNet and BERT model. The semantic knowledge extracted from side information is used to generate each image's sampling weight. In the training stage, we adopt the class balanced sampling strategy to handle the long-tail problem. For each class, we choose images with generated weights according to semantic knowledge to handle noise annotations.

Team: BigVideo

Modalities: Image, Query ID, text

1 Strong models: SEResNeXt152, OctaveResNet152, Res2Net152 etc.

2 Data filtering with NLP model: Use text data to filter out noisy images using BERT embedding

3 Training strategy: We perform expanded input sizes, de-noising, and model diversity through fine-tuning.

4 Ensemble strategy

Team: Huaweicloud

Modalities: Image, Query ID, meta information

Our work is implemented using Huawei MoXing framework [1], which slightly improve accuracy while being much faster in training. As for the algorithms, the main idea is to leverage the meta information of each image and from search engine to clean up the data, and knowledge distillation for handling noise labels, as well as heuristic algorithm for learning an ensemble model.

Team: Y_Y

Modilites: Image, Query ID

Architecture: ResNet, ResNext

Entry 1: 8 different models average vote, including resnext101, resnet101, which apply different sampling strategy

Entry 2: 8 different models weighted vote, including resnext101, resnet101, which apply different sampling strategy

Entry 3: 8 different models plus 3 retrieve results, weighted vote. Same models in entry 1 and entry 2.

Entry 4: 8 different models plus 12 other base models which mos

Team: PCI

Modalities: Image, Query ID

Architecture: ResNet101 and ResNet152

First we randomly select one million samples to train a course Resnet101 model, and we use this model to clean samples.

Second we use cleaned data to train Resnet101 and Resnet152 models separately.

Third we use all samples to finetune Resnet101 and Resnet152 models separately. At last we ensemble all of the models to get last result.