ACM multimedia

Seattle,USA Oct 12-16 2020

Webly Supervised Image Classification with Metadata: Automatic Noisy Label Correction via Visual-Semantic Graph

Jingkang Yang*, Weirong Chen*, Litong Feng,
Xiaopeng Yan, Huabin Zheng, Wayne Zhang

sensetime

RICE

香港中文大學
The Chinese University of Hong Kong

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Webly Supervised Image Classification

What?

- Utilizes online search engines to collect billions of web images and labels them with the query name (searching keyword)

Why?

- Human annotations are extremely time-consuming and expensive
- Can pre-train general vision models directly from large-scale web data

Webly Supervised Image Classification

Challenge: **Semantic label noise**

- A real-world problem that most images of a category deviate from its true semantic concept

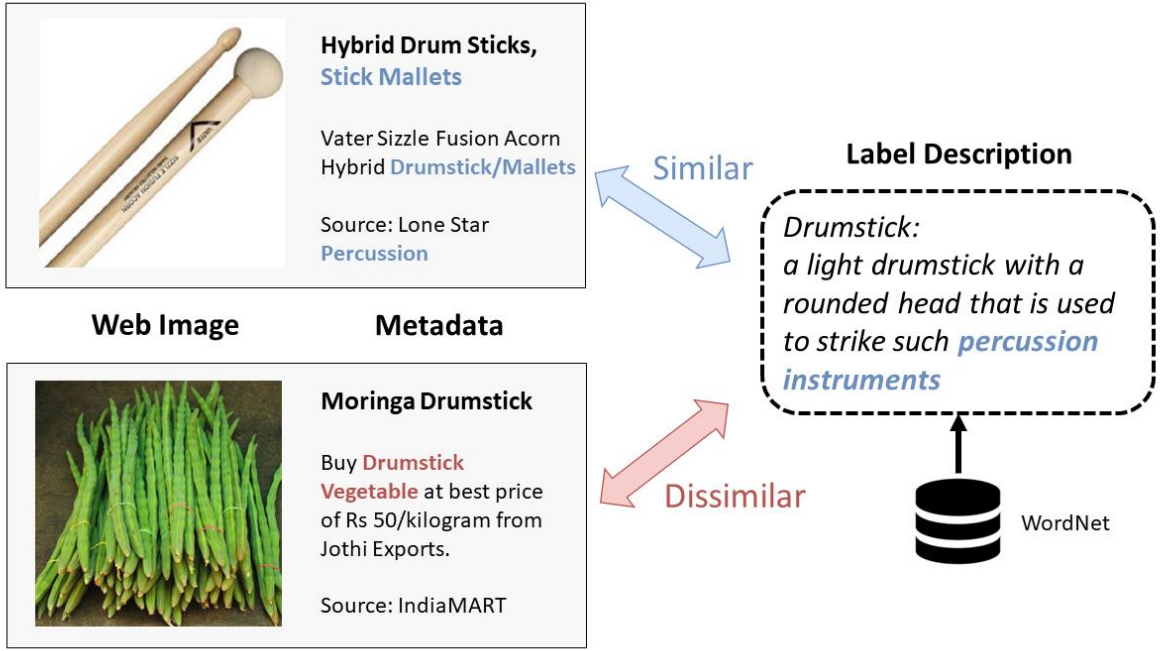
Query	Search Results	Correct Semantic
Tiger Cat (Compound)	   Tiger	 Tiger cat
Drumstick (Polysemy)	 Vegetable  Chicken Leg  Mallet	 Mallet

Two types of semantic confusion of query

Method

Insight 1: Metadata

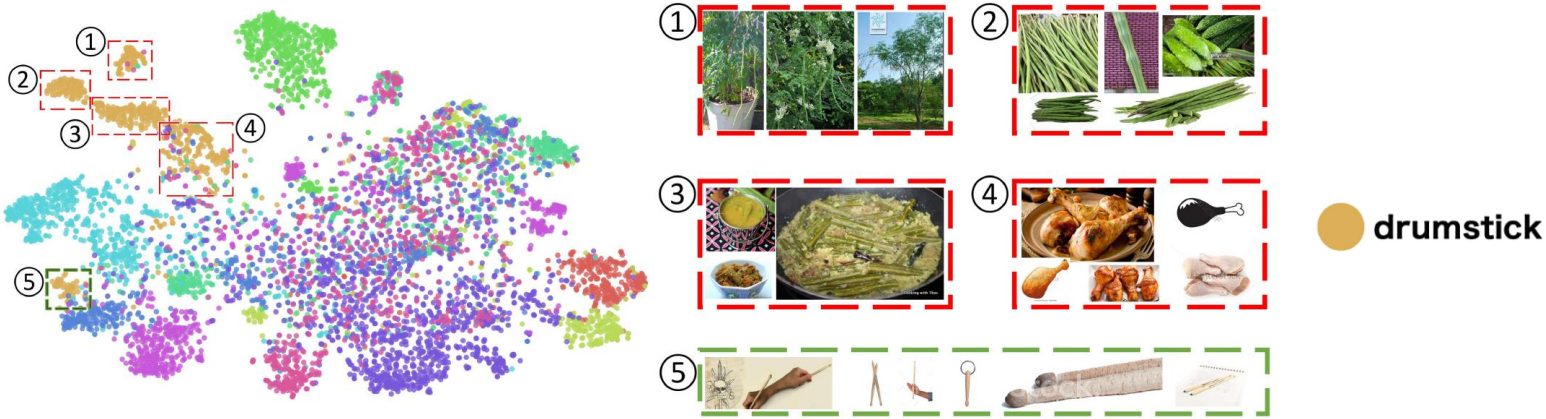
- Text metadata crawled along with web image can reflect image semantics
- Can handle severe semantic label noise problem automatically



Method

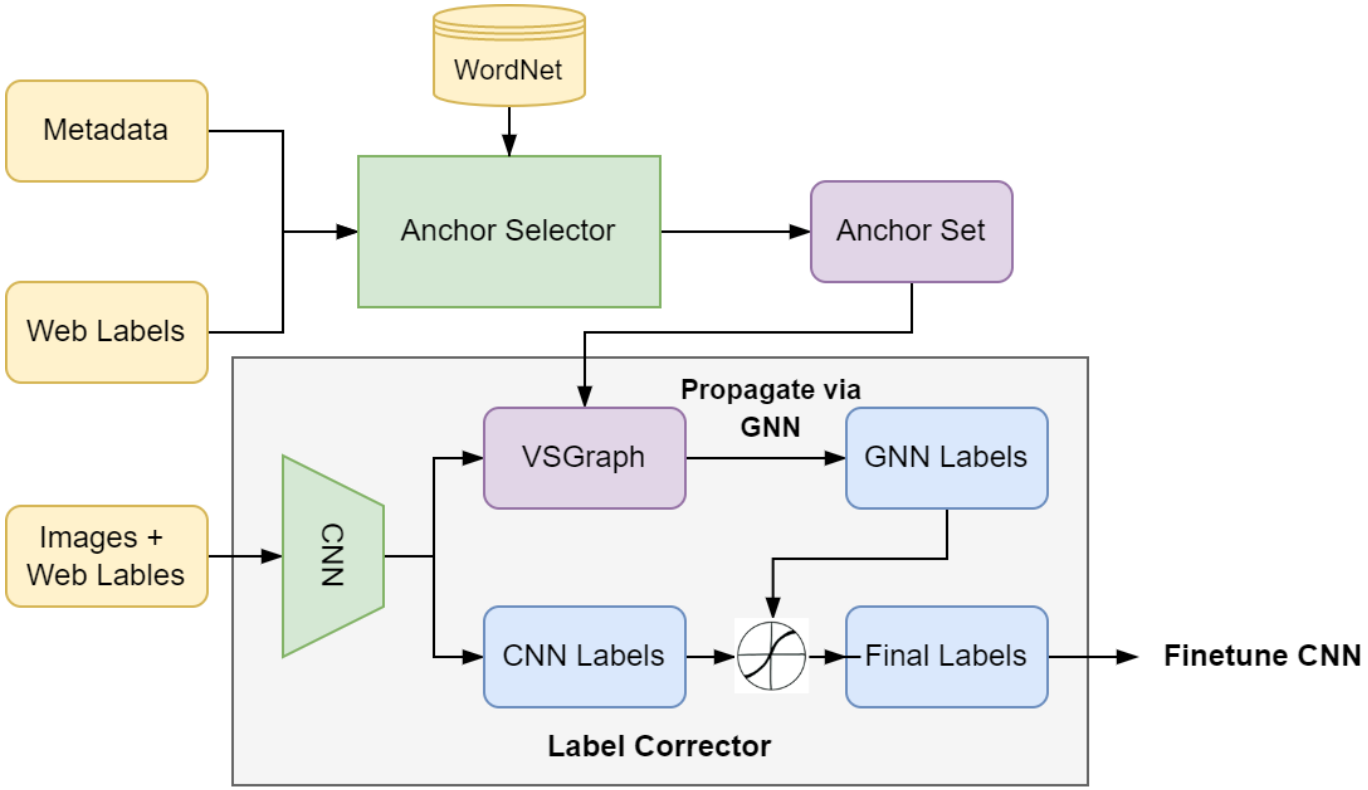
Insight 2: Visual-semantic Graph (VSGraph)

- Features that extracted from CNN models are clustered by semantics
- Clean samples can propagate correct semantic on VSGraph



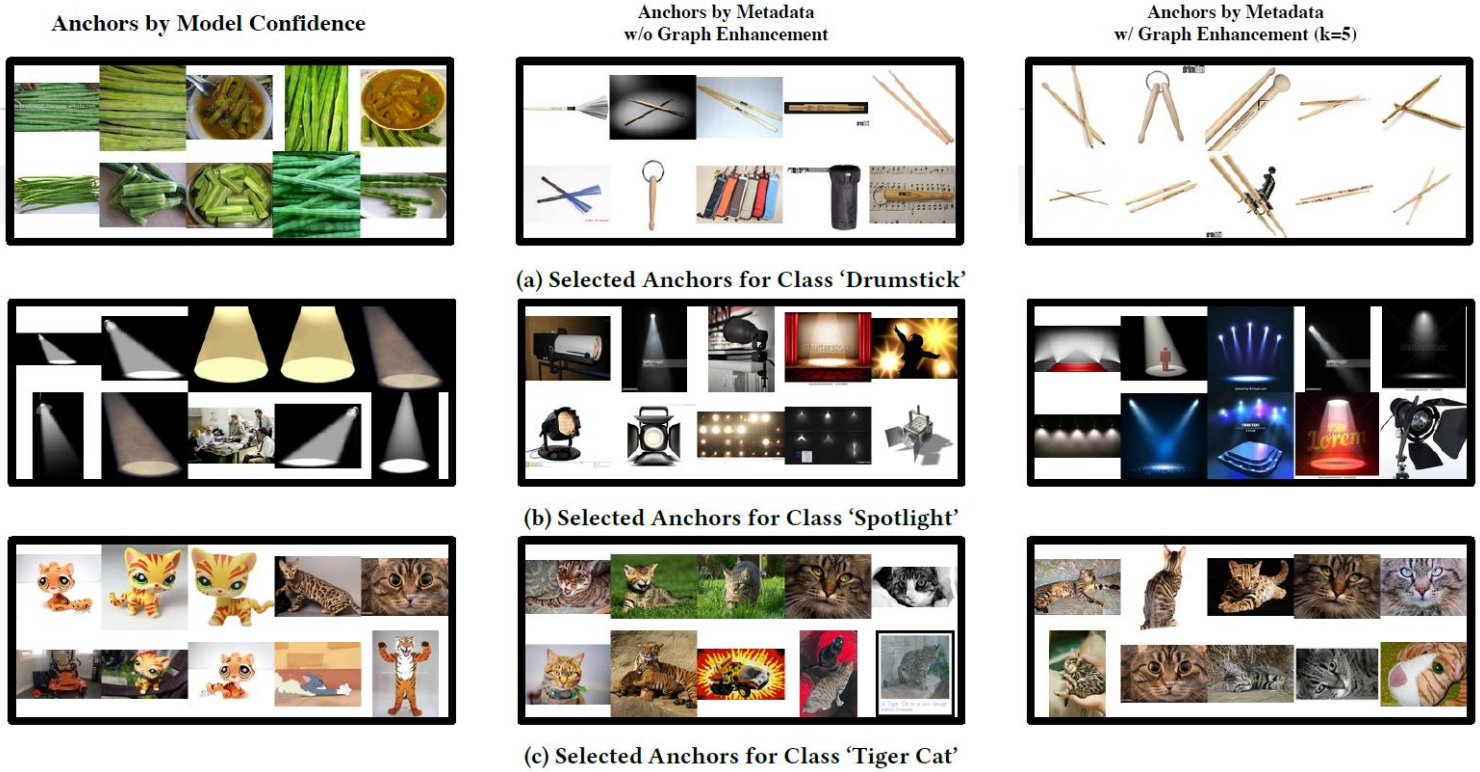
Web label 'Drumstick' shows representative images corresponding to 5 regions of interest. We observe that **similar semantics are clustered** and **different semantics are separated**.

Pipeline


WordNet
Metadata
Web Labels
Anchor Selector
Anchor Set
Propagate via GNN
VSGraph
GNN Labels
Images + Web Lables
CNN
CNN Labels
Final Labels
Finetune CNN
Label Corrector

Experimental Results

Performance: w/ Graph Enhancement > w/o Graph Enhancement > Model Confidence



(a) Selected Anchors for Class 'Drumstick'

(b) Selected Anchors for Class 'Spotlight'

(c) Selected Anchors for Class 'Tiger Cat'

Experimental Results

Our method achieves the **SOTA** performance on WebVision-1000

Table 2: The state-of-the-art results on WebVision-1000

Method	Backbone	WebVision		ImageNet	
		Top-1	Top-5	Top-1	Top-5
MentorNet [17]	InceptionResNetV2	72.60	88.90	64.20	84.80
CleanNet [24]	ResNet50	70.31	87.77	63.42	84.59
CurriculumNet [12]	InceptionV2	72.10	89.20	64.80	84.90
Multimodal [36]	InceptionV3	73.15	89.73	-	-
Pretrained model	ResNet50	74.25	89.84	68.28	86.23
Finetune by p_c only	ResNet50	75.15	89.93	69.07	86.76
Finetune by p_f	ResNet50	**75.48**	**90.15**	**69.42**	**87.29**

Table 3: Results on NUS-81-Web with noisy web labels for training. $K = 3$ is used for calculating C-F1 and O-F1

Method	C-F1	O-F1	mAP
Pretrained model	37.51	39.59	43.94
Finetune by p_c only	37.62	39.15	43.99
Finetune by p_f	**38.58**	**40.16**	**44.83**

Summary

- We highlight two understudied but critical factors in webly supervised learning: **semantic label noise** and **text metadata**
- **Visual Semantic Graph**: the webly pretrained CNN can provide reasonable visual feature space where similar images cluster themselves
- We design an effective and automatic label corrector by using clean anchor set with **GNN-based label propagation**