# Webly Supervised Concept Expansion for General Purpose Vision Models

Amita Kamath* Christopher Clark* Tanmay Gupta* Eric Kolve Derek Hoiem Aniruddha Kembhavi

Allen Institute for AI
University of Illinois Urbana-Champaign

ECCV TEL AVIV 2022

A cheap and effective way to expand your model's performance on various tasks across various concepts:

## Learn tasks from densely annotated data

COCO data

What sport is being played? *Baseball*

What is this? *Sports ball*

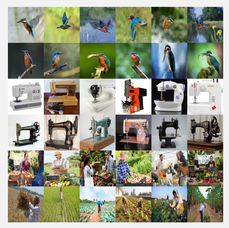
Locate the baseball glove.

Describe the image. *A little kid pitching a baseball at a game.*

Expensive, limited concepts

## Learn concepts from web search data

Web Search data



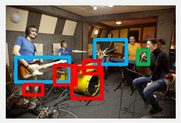
kingfisher flying
kingfisher sitting
electric sewing machine
vintage sewing machine
farmer selling
farmer harvesting

Cheap, many concepts

## Benefit from transfer!

What is the lady doing? Harvesting

Locate the guitar / drums / mic



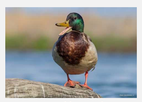
Caption the image. A person is working on a sewing machine

What is this?

kingfisher

bluejay

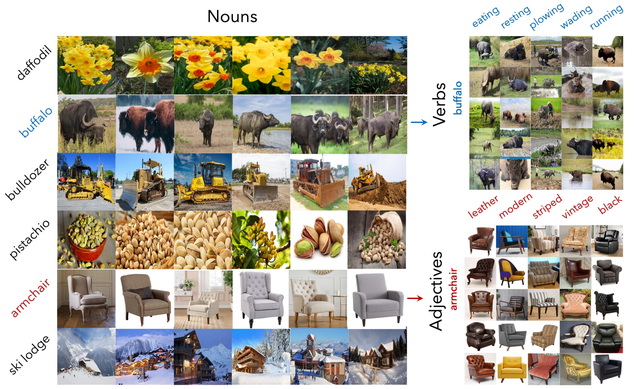
heron



mallard

## Why web data?

- High quality data for tail-end concepts
- Uncluttered, object-centric data
- Latest concepts, on-demand

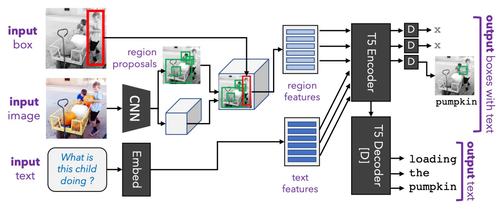



38K concepts (nouns/verbs/adjectives)
1M images
3.3M (templated) QA pairs

for only 154 USD!

e.g. "What is the buffalo doing?" "eating"

## Adding web data to training improves several models (including our new one) on data with unseen concepts!




### New Model: GPV-2

- Bounding box input
- Shared language decoder for all tasks (including localization!)
- Re-calibration to improve generalization

### New Benchmark: Diverse Concept Evaluation (DCE)

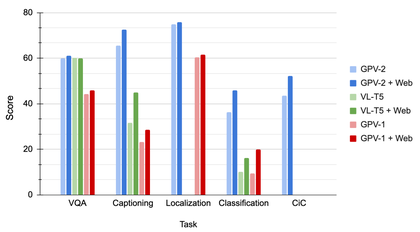
- ~500 concepts for 5 tasks
- Based on OpenImages, VisualGenome, and NoCaps
- Sampling improves balance in representation of various categories

Pre-cursor to GRIT:

- 7 vision and vision-language tasks
- Multiple data sources and diverse concepts per task
- Robustness and calibration evaluation

### Web data helps models perform tasks on more diverse concepts

- Evaluated 3 general-purpose vision-language models: GPV-2, GPV-1 [a], and VL-T5 [b]
- Adding web data to training improves generalization to new concepts, especially on captioning and classification




## Impressive performance!

| VQA | Captioning | Localization | Classification (cropped) | Classification in Context |
|---|---|---|---|---|
| What is the type of dress women wearing? | What is happening? | Find balance beam. | What is this thing? | What is this? |
| sari<br>scarves | a toddler wearing a hat riding a tricycle.<br>a small child in a hat riding a bike | solid lines<br>dotted lines | gondola<br>motorcycle | guacamole<br>broccoli |
| What color is the burrito? | Describe this image. | Locate cart in the image. | What object is this? | What is this object? |
| brown<br>green | a small blueberry muffin on a yellow plate.<br>a close up of a plate of food on a table | solid lines<br>dotted lines | harpsichord<br>suitcase | polar bear<br>sheep |

Legend: with web without web

## On-demand learning!

Ran 25 COVID-related concepts through our data collection pipeline, fine-tuned on the resulting data, and:

| VQA | Captioning | Localization | Classification (cropped) | Classification in Context |
|---|---|---|---|---|
| What is he holding? | Describe the image. | Find the temperature scanner. | What is this? | What is this? |
| covid vaccination card | a close up of a person wearing a kn95 mask | | nasal swab | pcr test |

[a] Gupta et al. "Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture" CVPR 2022
[b] Cho et al. "Unifying Vision-and-Language Tasks via Text Generation" ICML 2021