Computer Vision in the Wild: Benchmark & Challenge Summary

October 2022

Chunyuan Li
Deep Learning Team
Microsoft Research, Redmond
Why CVinW? Evaluation of Language-augmented Visual Task-level Transfer

Current Research (in the lab) → Computer Vision in the Wild

Image Classification
- ImageNet
- Food
- Flowers
- Textures
- Facial Expression
- Traffic Signs

Object Detection
- COCO/LVIS
- Mask Wearing
- Thermal
- Potholes
- Wildfire Smoke
- Paper-Rock-Scissors

Trend
- Building transferable systems (e.g., foundation models) that can adapt to a wide range of CV tasks
- Inspired by the success of CLIP, many language-augmented visual models appear

Challenges
- **Fairness**: Customized task sets may favor individual pre-trained model
- **Transparency**: Detailed model adaptation process is inaccessible
What is Computer Vision in the Wild (CVinW)?

Developing a transferable foundation model/system that can *effortlessly* adapt to a *large range of visual tasks* in the wild.

It comes with two key factors:

1. The task transfer scenarios are broad
2. The task transfer cost is low.

CVinW vs other CV settings

Open-Set

Concept / Vocabulary / Language

border collie, running, while shirt

mask-wearing food flowers textures

Close-Set

Training Domain

Domain Shift

Visual Content

Computer Vision In the Wild

person, sheep, dog

person, dog

person, dog, standing/sitting
2D space for the definition of adaptation cost

Parameter-Efficiency

Fine-Tuning

Linear Probing

Prompt with Frozen Models

Sample-Efficiency

Zero-shot  Few-shot  Full-shot

Most inexpensive

Most expensive

Adaptation Cost

Low  High
Where to start?

**ELEVATER:**
A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models

https://arxiv.org/abs/2204.08790  NeurIPS 2022 (Benchmarks and Datasets Track)

Chunyuan Li\(^1\), Haotian Liu\(^2\), Liunian Harold Li\(^3\), Pengchuan Zhang\(^1\), Jyoti Aneja\(^1\)
Jianwei Yang\(^1\), Ping Jin\(^1\), Houdong Hu\(^1\), Zicheng Liu\(^1\), Yong Jae Lee\(^2\), Jianfeng Gao\(^1\)
\(^1\)Microsoft  \(^2\)University of Wisconsin–Madison  \(^3\)UCLA
Benchmarks: **ELEVATER**

- **Dataset Suite**
  - Image Classification: **20 datasets**
  - Object Detection: **35 datasets**

**Objects**
- **Flowers102**
- **DTD**
- **Food101**
- **RESISC45**
- **Country211**
- **Caltech101**
- **FGVC**
- **Aircraft**
- **EuroSat**
- **VOC2007**
- **StanfordCars**
- **MNIST**
- **GTSRB**
- **OxfordPets**
- **CIFAR100**
- **CIFAR10**

**Categories**
- ChessPieces
- BrackishUnderwater
- PascalVOC
- AerialMaritimeDrone
- Large
- Small
- OpenPoetry
- Vision
- Website
- Screenshots
- Plant
- doc
- Aquarium
- Dice
- Bogle
- Boards
- American
- SignLanguage
- Letters
- Uno
- Cards
- Vehicles
- OpenImages
- Drone
- Control
- SelfDriving
- Car
- OxfordPets
- breed
- EgoHands
- specific
- Aerial
- Maritime
- Drone
- tiled
- Egohands
- general

- **External Knowledge**
  - WordNet, Wiktionary, GPT-3

- **Concept name**: risotto
  - **Def_wik**: An Italian savoury dish made with rice and other ingredients
  - **Def_wn**: rice cooked with broth and sprinkled with grated cheese
  - **Path_wn**: [risotto, dish, nutriment, food, substance, matter, physical_entity, entity]
  - **GPT3**: A rice dish made with arborio rice and typically served with meat or fish
Benchmarks: A more diverse set of tasks

ImageNet → Image Classification in the Wild (20 datasets)

LVIS → Object Detection in the Wild (35 datasets)

Semantic Space using PCA

Task diversity using std

0.610 0.680

0.533 0.619
Where to submit results?  Challenge → Track → Phase

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Track</th>
<th>Definition (*IN = ImageNet )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>No IN-1K data; Scaling Success</td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>No IN-1K data; Limited pretrain data (IN21K, CC3M+12M, YFCC15M)</td>
<td></td>
</tr>
<tr>
<td>ICinW</td>
<td>IN1K is allowed in pretraining, eg, self-supervised Learning</td>
<td></td>
</tr>
<tr>
<td>Parameter-Efficiency</td>
<td>Efficient model adaption methods</td>
<td></td>
</tr>
<tr>
<td>ODinW</td>
<td>Zero-Shot</td>
<td>No Training examples in ODinW are used</td>
</tr>
<tr>
<td></td>
<td>Full-shot</td>
<td>All Training examples in ODinW are used</td>
</tr>
</tbody>
</table>
A collaborative community-effort
-- to benchmark the SOTA foundation vision models
Challenge Talks
-- one talk for each track

- ICinW Industry Track | Chinese CLIP | Junyang Lin (Alibaba)
- ICinW Academic Track | K-LITE | Sheng Shen (University of California, Berkeley)
- ICinW ImageNet-1K in Pre-training | Bamboo | Yuanhan Zhang (Nanyang Technological University)
- ICinW Parameter-Efficiency | ProDA | Yuning Lu (University of Science and Technology of China)
- ODinW Zero-Shot Track | DetCLIP | Jianhua Han (Huawei)
- ODinW Full-Shot Track | DINO | Shilong Liu (IDEA & Tsinghua)

Criterion

• **Ranking:** Top ranked methods by October 20, 2022 (3 days ago)

• **Availability:** Make a video presentation before Challenge

• **Deduplication:** No duplicated presentation between workshop paper and challenge

• **Note:** The ranking has been changing in the past 3 days
• Larger models are better (4B > 1.6B > 1.0B > 0.4B)

• Foreign image-text pre-training is effective to IC in English

• Generative models such as GIT has a large space to improve for IC

• ALIGN is comparable with CLIP
- MaskCLIP ranks 1st on ICinW
- External knowledge is useful
- The conclusions are inconsistent between ICinW and ImageNet-1K; be more careful when designing architectures and objectives
- Not all models outperform CLIP trained on YFCC
• Bamboo ranks 1\textsuperscript{st} on ICinW; Data-centric AI is effective

• Image self-supervised learning methods are popular

• Zero-shot FLAVA outperforms FT and LP many models

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Method</th>
<th>Average Score</th>
<th># Vision Backbone Params [M]</th>
<th># Trainable Params [K]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bamboo</td>
<td>Bamboo-ViTB/16 LP</td>
<td>63.7</td>
<td>85.8</td>
<td>34.6</td>
</tr>
<tr>
<td>2</td>
<td>ELEVATER</td>
<td>ViT [ViT-B/16, LP]</td>
<td>57.6</td>
<td>85.8</td>
<td>44.3</td>
</tr>
<tr>
<td>3</td>
<td>ELEVATER</td>
<td>ViT [ViT-B/16, FT]</td>
<td>57.2</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>4</td>
<td>Amazon-m5</td>
<td>DeCL [ViT-B/16, LP]</td>
<td>54.7</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>5</td>
<td>CACR</td>
<td>CACR [ViT-B/16, LP]</td>
<td>54.5</td>
<td>85.8</td>
<td>7.69</td>
</tr>
<tr>
<td>6</td>
<td>ELEVATER</td>
<td>DeiT [ViT-B/16, LP]</td>
<td>54.1</td>
<td>85.8</td>
<td>44.3</td>
</tr>
<tr>
<td>7</td>
<td>ELEVATER</td>
<td>DeiT [ViT-B/16, FT]</td>
<td>54.1</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>8</td>
<td>CACR</td>
<td>CACR [ViT-B/16, LP]</td>
<td>53.6</td>
<td>85.8</td>
<td>7.69</td>
</tr>
<tr>
<td>9</td>
<td>ELEVATER</td>
<td>MoCo-v3 [ViT-B/16, LP]</td>
<td>50.2</td>
<td>85.8</td>
<td>44.3</td>
</tr>
<tr>
<td>10</td>
<td>CACR</td>
<td>CACR [ViT-B/16, FT]</td>
<td>48.9</td>
<td>85.8</td>
<td>86424.05</td>
</tr>
<tr>
<td>11</td>
<td>Facebook AI Research</td>
<td>FLAVA (PMD-ZeroShot)</td>
<td>48.7</td>
<td>241.36</td>
<td>0.0</td>
</tr>
<tr>
<td>12</td>
<td>Bamboo</td>
<td>Bamboo-ViTB/16 FT</td>
<td>48.4</td>
<td>85.8</td>
<td>85833.26</td>
</tr>
<tr>
<td>13</td>
<td>SupMAE</td>
<td>SupMAE [ViT-B/16, FT]</td>
<td>46.8</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>14</td>
<td>CAE v1</td>
<td>Baidu (CAE v1 [ViT-B/16, LP])</td>
<td>44.1</td>
<td>85.81</td>
<td>44.26</td>
</tr>
<tr>
<td>15</td>
<td>ELEVATER</td>
<td>MoCo-v3 [ViT-B/16, FT]</td>
<td>39.3</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>16</td>
<td>CAE v1</td>
<td>Baidu (CAE v1 [ViT-B/16, FT])</td>
<td>37.9</td>
<td>85.81</td>
<td>85807.87</td>
</tr>
<tr>
<td>17</td>
<td>ELEVATER</td>
<td>MAE [ViT-B/16, FT]</td>
<td>36.1</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>18</td>
<td>ELEVATER</td>
<td>MAE [ViT-B/16, LP]</td>
<td>33.4</td>
<td>85.8</td>
<td>44.3</td>
</tr>
<tr>
<td>19</td>
<td>ELEVATER</td>
<td>From-Scratch [ViT-B/16, FT]</td>
<td>20.8</td>
<td>85.8</td>
<td>85842.96</td>
</tr>
<tr>
<td>20</td>
<td>CV-Team</td>
<td>ViT-Contrastive</td>
<td>20.7</td>
<td>85.8</td>
<td>86570.73</td>
</tr>
<tr>
<td>21</td>
<td>ELEVATER</td>
<td>From-Scratch [ViT-B/16, LP]</td>
<td>19.6</td>
<td>85.8</td>
<td>44.3</td>
</tr>
</tbody>
</table>
A single number that measures both prediction accuracy and parameter-efficiency

\[ PE = \text{score} \times \exp(\log_{10}(\# \text{ trainable-parameters}/M_0 + 1)) \]

- ProDA ranks the 1\textsuperscript{st}
- Advanced adaptation methods from NLP are not really better than linear probing?
• Florence ranks 1st with average;
• DetCLIP ranks 1st with median

• Larger pre-training dataset leads to better performance; Though MDETR trained on a small dataset, the results are not bad

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team</th>
<th>Method</th>
<th>Average Score</th>
<th>Median Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Florence</td>
<td>FL-1.5-D5</td>
<td>25.8</td>
<td>14.3</td>
</tr>
<tr>
<td>2</td>
<td>DetCLIP-team</td>
<td>DetCLIP</td>
<td>24.9</td>
<td>18.3</td>
</tr>
<tr>
<td>3</td>
<td>GLIPv2_team</td>
<td>GLIPv2-T</td>
<td>22.3</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td>OmLab</td>
<td>OmDet</td>
<td>19.7</td>
<td>8.9</td>
</tr>
<tr>
<td>5</td>
<td>ODinW_Team</td>
<td>GLIP-T</td>
<td>19.6</td>
<td>5.1</td>
</tr>
<tr>
<td>6</td>
<td>FIBER</td>
<td>FIBER</td>
<td>19.5</td>
<td>10.4</td>
</tr>
<tr>
<td>7</td>
<td>OmLab</td>
<td>OmDet</td>
<td>19.0</td>
<td>8.9</td>
</tr>
<tr>
<td>8</td>
<td>Google Research</td>
<td>OWL-ViT L/14 @ 672</td>
<td>18.8</td>
<td>9.8</td>
</tr>
<tr>
<td>9</td>
<td>ODinW_Team</td>
<td>GLIP-T (B)</td>
<td>12.8</td>
<td>2.2</td>
</tr>
<tr>
<td>10</td>
<td>ODinW_Team</td>
<td>GLIP-T (A)</td>
<td>11.4</td>
<td>1.6</td>
</tr>
<tr>
<td>11</td>
<td>MDETR-NYU</td>
<td>MDETR - ENB5</td>
<td>10.7</td>
<td>3.0</td>
</tr>
<tr>
<td>12</td>
<td>MDETR-NYU</td>
<td>MDETR - ENB3</td>
<td>10.1</td>
<td>2.7</td>
</tr>
<tr>
<td>13</td>
<td>MDETR-NYU</td>
<td>MDETR - R101</td>
<td>9.9</td>
<td>3.1</td>
</tr>
</tbody>
</table>
OmLab ranks 1st with average and median;

DINO, the best performing OD head on COCO, performs well on ODinW

Results on October 22, 2022 PT
Call for Collaboration
-- benchmarking the transfer ability of SoTA vision models

• **Criterion**
  • **Goals:**
    • Summary of Challenge results on ICinW and ODinW
    • A comprehensive technical report to benchmark the best vision checkpoints and adaptation methods
    • A shared view to push CVinW
  • **Authorship:** Contributors with valid submissions are encouraged to co-author the report
  • **Timeline:** The 1st version by the end of 2022; Continual updating arXiv when necessary

• **Inspiring Examples**  BIGBench and BigScience in NLP

• **Future Update**
# Challenge Presentation Agenda

<table>
<thead>
<tr>
<th>Talk</th>
<th>Challenge</th>
<th>Track</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>iCinW</td>
<td>Industry</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Academic</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>ImageNet-1K in Pretraining</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Parameter-Efficiency</td>
</tr>
<tr>
<td>5</td>
<td>ODinW</td>
<td>Zero-Shot</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Full-shot</td>
</tr>
</tbody>
</table>

CVinW
Computer Vision in the Wild