Tools Identification By On-Board Adaptation of Vision-and-Language Models

Jun Hu, Phil Miller, Michael Lomnitz, Saurabh Farkya, Emre Yilmaz, Aswin Raghavan, David Zhang, Michael Piacentino

SRI International
aswin.raghavan@sri.com

Abstract

A robotic workshop assistant has been a long-standing grand challenge for robotics, speech, computer vision, and artificial intelligence (AI) research. We revisit the goal of visual identification of tools from human queries in the current era of Large Vision-and-Language models (like GPT-4 (OpenAI 2023)). We find that current off-the-shelf models (that are trained on internet images) are unable to overcome the domain shift and unable to identify small, obscure tools in cluttered environments. Furthermore, these models are unable to match tools to their intended purpose or affordances. We present a novel system for online domain adaptation that can be run directly on a small on-board processor. The system uses Hyperdimensional Computing (HD) (Kanerva 2009), a fast and efficient neuromorphic method. We adapted CLIP (Radford et al. 2021) to work with explicit ("I need the hammer") and implicit purpose-driven queries ("Drive these nails"), and even with depth images as input. This demo allows the user to try out various real tools and interact via free-form audio.

System Demonstration

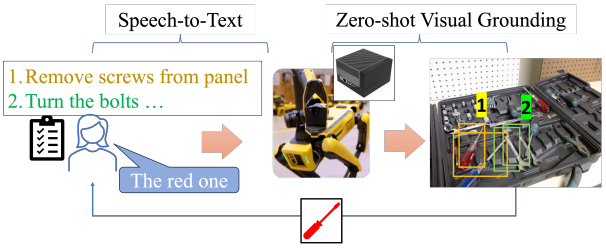

Figure 1: Example workflow speech-to-text and visual grounding. Robot fetching the tool will not be demonstrated.

The setup for the demonstration is shown in Figure 2. Visual input comes from a webcam mounted on a tripod that can be moved around the display area. A headset with a microphone is used for audio. A set of tools and pegboard will be provided. The adapted models will be demonstrated on laptops and the NVIDIA Jetson Orin (Barnell et al. 2022). New tools can be enrolled during the demo. The identified tools will be displayed on monitors. Users will be able to modify components and interact with different queries.

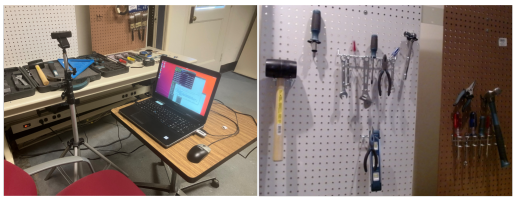

Figure 2: Demo setup: Webcam, Orin, tools and pegboard.

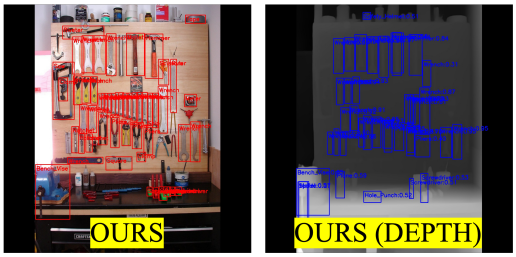

Figure 3: Example results showing precision of our method.

Our method for domain adaptation is a two-stage pipeline. The first stage is a region proposal network (RPN) (Ren et al. 2015) that outputs thousands of boxes. The RPN can be agnostic to the domain or classes. The second stage is a classifier that maps each box into a category or as background.

In order to improve the generalization between domains, we propose to incorporate text (from audio) as an additional input to the classifier, e.g. combining the appearance of objects (e.g., say a shovel) with its purpose (e.g., to dig a hole) can help generalize to new tools with similar appearance on a part of the tool (e.g., a spade can also dig a hole). We use a vision and language model called CLIP (Radford et al. 2021). CLIP can embed images and text in a vector space such that an image and its caption will have high cosine similarity. Our method applies CLIP to each proposal (the cropped image). Each proposal is captioned with either the name of the tool, or the purpose of the tool (Table 1), or "background" for ROIs with insufficient overlap with ground truth boxes.

We use Hyperdimensional Computing (HD) (Kanerva

Tool	Purpose 1	Purpose 2
Hammer	drive nails	remove nails
Level	check horizontal	check vertical
Square	measure 90-degree	measure length

Table 1: Example tool and purpose captions.

Algorithm 1: Training / Inference in Target Domain

Input: RGB or depth images $\mathcal{I}$
Input: (In training) Labelled bounding boxes $\mathcal{B}$
Parameter: Size of HD vectors D
Output: (Training) HD Exemplars E for source categories

Initialize training
$e \leftarrow \mathbf{0}^D$ for each $e \in E$
# Loop if training
for each training image I **do**
  Boxes $B \leftarrow$ RegionProposals(I)
  for each box $b \in B$ **do**
    $\phi_b \leftarrow$ CLIP(b); $v_b \leftarrow$ EncodeHD(ϕ_b)
    # In Training (inference caption comes from user)
    caption $\leftarrow b$.label from $\mathcal{B}$ else "background"
    $\phi_Q \leftarrow$ CLIP(caption); $v_Q \leftarrow$ EncodeHD(ϕ_Q)
    $V \leftarrow v_q \oplus v_b$
    # In Training
    E[label] $\leftarrow E$[label] + V; n[label] $\leftarrow n$[label] + 1
    # In inference
    $d \leftarrow \frac{1}{D}||V - e||_1$ for each exemplar $e \in E$
    Label b with caption if $\max(Softmax(\text{1-d})) > \epsilon$
  end for
end for
# In training
$E[i] \leftarrow \mathbf{1}_D(\frac{E[i]}{n[i]} > 0.5)$ for each label i
return E

2009) to perform online domain adaptation (Alg. 1). We randomly project embedding vectors (from CLIP) to a high-dimensional binary vector space ("EmbedHD" function). We apply the same projection to visual and text embedding, preserving any semantic similarities. The visual and text HD vectors are combined using XOR ("binding") and average ("bundling") to produce exemplars E. We form two exemplars per tool using names and purpose captions. During inference, we need to handle arbitrary queries and new tools. Inference treats each ROI and query pair as a binary classification problem: given the user's query and ROI, the probability that the ROI is not captioned by the query is the hamming distance to exemplars E.

Zero-shot mAP	Vision	Vision+Language
ALET to SKIML	9.5	20.5
SKIML to ALET	5	20

Table 2: Impact of vision+language on zero-shot mAP.

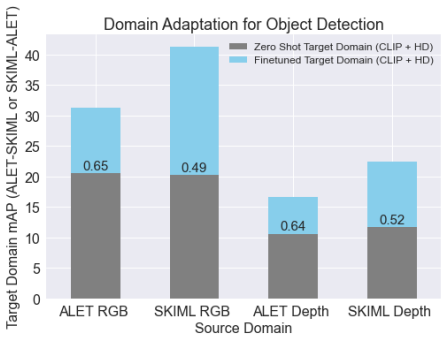

Figure 4: Mean Average Precision (mAP) for zero-shot (gray) and finetuned (blue) object detection after training on source domain (x-axis) and tested on target domain (y-axis).

Eval. purpose queries	mAP@0.5	@0.5:0.95
Exemplars from tool name	27.61	18.43
Exemplars from purposes	43.57	28.83

Table 3: Impact of exemplars from purpose annotations.

Results

We use two datasets for evaluating domain adaptation. The ALET dataset (Kurnaz et al. 2020) consists of 50 tools with multiple tools per image in cluttered environments. The SKIML dataset is a proprietary dataset with 12 tools with one tool per image. SKIML and ALET classes have some common tools but not all. We train on SKIML and test on ALET and vice-versa to evaluate zero-shot performance. Figure 4 shows Mean Average Precision (mAP) for object detection. Our method achieves non-trivial zero-shot mAP that is about 50% of the finetuned mAP on the target domain and able to transfer from 12 SKIML tools to 50 ALET tools (20 mAP zero-shot vs 40 mAP finetuning). The high zero-shot precision is due to the combination of vision and language modalities (Table 2). Table 3 shows incorporating purpose annotations via new exemplars improves precision.

Automatic speech-to-text uses a wav2vec2 (Baevski et al. 2020) architecture. We finetuned this model on our in-house English training data which contains 2250 hours of speech. Our system was tested on the edge device, NVIDIA Orin, and validated against a GPU server (NVIDIA A5000). A comparison of the time to process on image is shown in Table 4. The time is dominated by generating region proposals. However, proposals have to be generated only once to process all queries. When a new query is entered, only the CLIP text embedding and the hamming distances have to be recomputed and the binary HD operations are fast.

Runtime (s/image)	Region Proposals	CLIP	HD
A5000 GPU	4.38	2.30	0.52
NVIDIA Orin	18.08	6.85	1.95

Table 4: Runtime on server vs Orin edge device.

References

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477.

Barnell, M.; Raymond, C.; Smiley, S.; Isereau, D.; and Brown, D. 2022. Ultra Low-Power Deep Learning Applications at the Edge with Jetson Orin AGX Hardware. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*.

Kanerva, P. 2009. Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. *Cognitive Computation*.

Kurnaz, F. C.; Hocaoğlu, B.; Yılmaz, M. K.; İdil Sülo; and Kalkan, S. 2020. ALET (Automated Labeling of Equipment and Tools): A Dataset, a Baseline and a Usecase for Tool Detection in the Wild. arXiv:1910.11713.

OpenAI. 2023. GPT-4 Technical Report. Technical report.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

This research is based upon work supported in part by the Office of the Intelligence Advanced Research Projects Activity (IARPA) via Contract No: 2022-21100600001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein