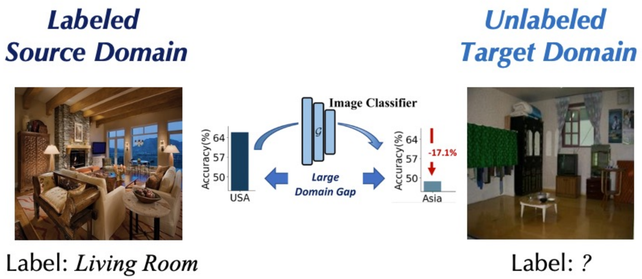
Tell, Don`t Show!: Language Guidance Eases Transfer Across Domains in Images and Videos

Tarun Kalluri *Bodhisattwa Prasad Majumder* *Manmohan Chandraker*

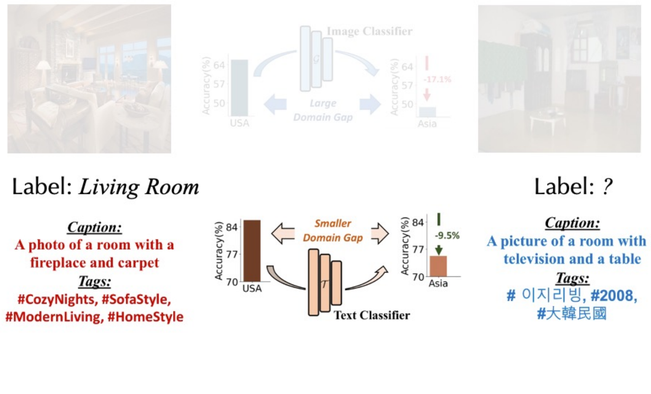
ICML International Conference On Machine Learning

Lesser Domain Shifts in Text Modality

- Unsupervised domain adaptation in pixel space is often difficult for large shifts.



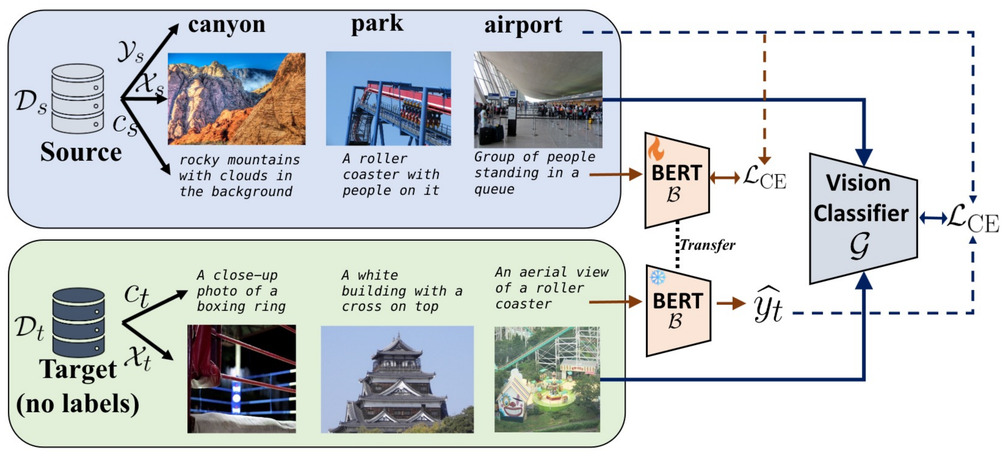
- The corresponding text-captions, however, have much lesser cross-domain drops with richer semantics.



Got only few seconds? Here's a tldr!

When faced with significant distribution shift in images and videos, train a robust text classifier using the captions instead, and distill its predictions into the vision classifier.

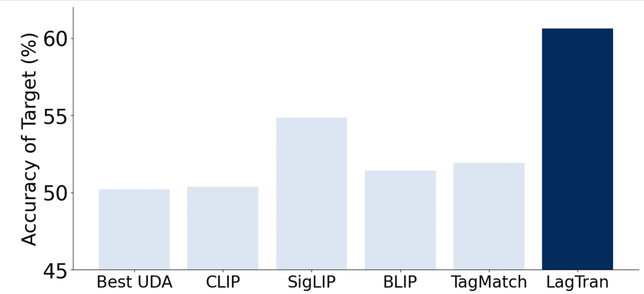
Visual Supervision Using Language Guided Pseudo-labels



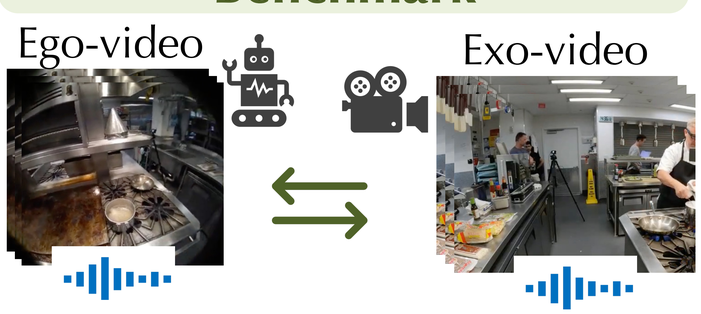
We improve transfer by:

1. Training a text-classifier using the source domain captions.
2. Predicting pseudo-labels for target domain captions.
3. Using these pseudo-labels for supervised training target domain images.

Best Accuracy on Challenging GeoNet Data...



... and a New Ego2Exo Benchmark



- Benchmark to study transfer between ego and exo videos.
- >10k videos, 24 labels, atomic narrations for all videos.