UC San Diego

ICCV23 PARIS

Distilling Large Vision-Language Model with Out-of-Distribution Generalizability

Xuanlin Li*, Yunhao Fang*, Minghua Liu, Zhan Ling, Zhuowen Tu, Hao Su

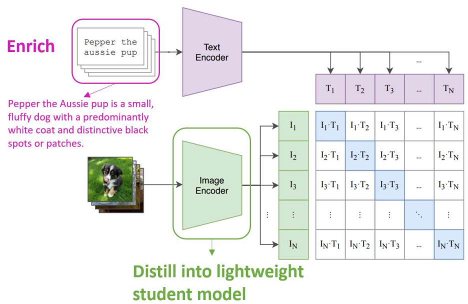
University of California San Diego

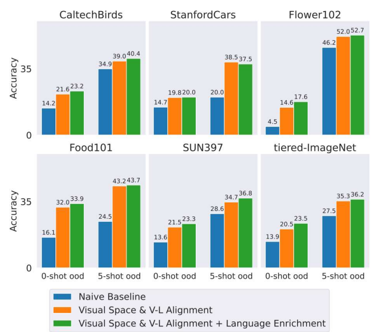
Special Thanks: TILOS, Qualcomm AI research

Paper Link

Overview

- **Goal**: **Distill visual representations in large vision-language teachers into lightweight student models,** while allowing students to possess **strong open-vocabulary generalization ability towards out-of-distribution (OOD) concepts**.
  - Potential for deployment on mobile / IoT devices and robotics scenarios.
  - Experiments are conducted on small-to-medium datasets, allowing for fast research & development (R&D) cycles and is practical for lower-data regimes like 3D and robotics.
- **Main Findings: 2 principles to enhance student's OOD generalizability:**
  - **Better imitate teacher's visual representation space**, and carefully promote **better vision-language alignment coherence** with the teacher.
  - **Enrich teacher's language representations** with more finegrained & meaningful attributes to effectively distinguish between different labels, both during distillation and inference. **This can be accomplished by prompting LLMs like ChatGPT.**





Teacher-Student Visual Space Alignments

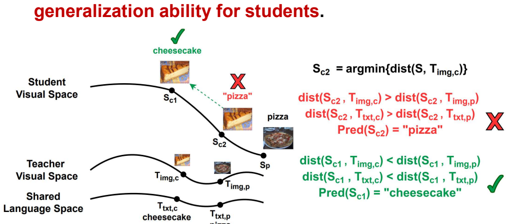
- **Precisely matching teacher and student's high-dimensional visual spaces is inherently challenging** (with high MSE feature matching loss).
- In this case, **minimizing student-teacher visual feature distance does NOT yield the best OOD generalization ability for students.**

$$\mathcal{L}_{cls}(\mathbf{x}, y) = \sum_{y'} -\mathbf{1}_{y'=y} \log P_S(y'|\mathbf{x})$$

$$P_S(y|\mathbf{x}) = \frac{\exp(\cos(S(\mathbf{x}), T_{txt}(l(y)))/\tau)}{\sum_{y' \in \mathcal{Y}} \exp(\cos(S(\mathbf{x}), T_{txt}(l(y')))/\tau)}$$

$$\mathcal{L}_{mse}(\mathbf{x}) = ||S(\mathbf{x}) - T_{img}(\mathbf{x})||_2^2$$

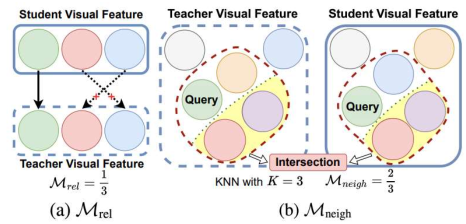
$$\mathcal{L}_{im\text{-}cst}(\mathbf{x}) = \frac{\exp(-||S(\mathbf{x}) - T_{img}(\mathbf{x})||_2^2/\tau)}{\sum_{\mathbf{x}'} \exp(-||S(\mathbf{x}) - T_{img}(\mathbf{x}')||_2^2/\tau)}$$



	Food101	SUN397
$\mathcal{L}_{mse}$	0.24 / 28.4°	0.36 / 34.9°
$\mathcal{L}_{mse}$ (RN50)	0.24 / 28.4°	0.35 / 34.4°
$\mathcal{L}_{cls} + \mathcal{L}_{mse}$	0.45 / 39.2°	0.71 / 49.8°
$\mathcal{L}_{cls} + \mathcal{L}_{mse} + \mathcal{L}_{im\text{-}cst}$	0.65 / 47.6°	0.82 / 53.8°
$\mathcal{L}_{cls} + \mathcal{L}_{im\text{-}cst}$	1.29 / 69.2°	1.28 / 68.9°

Average MSE / degree difference between student & teacher visual features for students trained with different strategies. Student: RN18; Teacher: CLIP ViT-L/14

- However, teacher-student coherence in **local visual space structures** and **relative visual feature relationships** is crucial for student OOD generalization, as they *implicitly* enhance vision-language alignments.
- We find that **Contrastive losses** like L_{im-cst} are especially helpful for student OOD generalization, as it **enables much better student-teacher proximity in both local and relative visual feature structures.**



$$\mathcal{M}_{rel}(\mathcal{X}) = \frac{\sum_{i=1}^{|\mathcal{X}|} \mathbf{1}[i = \mathrm{argmin}_j ||T_{img}(\mathbf{x}_j) - S(\mathbf{x}_i)||_2^2]}{|\mathcal{X}|}$$

$$\mathcal{M}_{neigh}(\mathcal{X}, k) = \frac{\sum_{i=1}^{|\mathcal{X}|} |\mathrm{KNN}(S, \mathbf{x}_i, k) \cap \mathrm{KNN}(T_{img}, \mathbf{x}_i, k)|}{k|\mathcal{X}|}$$

$\mathcal{M}_{rel} \uparrow$ | $\mathcal{M}_{neigh} \uparrow$	$\mathcal{X}_{train}$	$\mathcal{X}_{ood}$	$k = 3$	$k = 5$	$k = 10$
$\mathcal{L}_{cls} + \mathcal{L}_{mse}$	0.030	0.004	0.13 / 0.06	0.18 / 0.07	0.27 / 0.08
$\mathcal{L}_{cls} + \mathcal{L}_{mse} + \mathcal{L}_{im\text{-}cst}$	0.305	0.022	0.20 / 0.10	0.25 / 0.11	0.34 / 0.13

Teacher-Student Vision-Language Alignments

- *Explicitly* enhance teacher-student vision-language alignment coherency.
- **Carefully** preserve teacher's vision-language alignment structure **by accounting for teacher misalignments** (through e.g., L_{vlprox}).

$$\mathcal{L}_{vlprox}(\mathbf{x}, k) = I(\mathbf{x}) \cdot \mathcal{D}_{KL}(P_{T,topk}(\cdot|\mathbf{x}) || P_{S,topk}(\cdot|\mathbf{x}))$$

$$I(\mathbf{x}) = \mathbf{1}[\mathrm{argmax}_y P_T(y|\mathbf{x}) = \mathrm{label}(\mathbf{x})]$$

$$P_{\cdot,topk}(y|\mathbf{x}) = \frac{\mathbf{1}_{y \in Y_{topk}} P_\cdot(y|\mathbf{x})}{\sum_{y \in Y_{topk}} P_\cdot(y|\mathbf{x})}; Y_{topk} = \mathrm{argtopk}_y P_T(y|\mathbf{x})$$

$\mathcal{M}_{vlalign} \downarrow$	$k = 2$	$k = 3$	$k = 5$
$\mathcal{L}_{cls} + \mathcal{L}_{mse}$	0.20 / 0.50	0.68 / 1.45	2.67 / 4.73
$\mathcal{L}_{cls} + \mathcal{L}_{mse} + \mathcal{L}_{im\text{-}cst}$	0.18 / 0.43	0.62 / 1.3	2.52 / 4.24
$\mathcal{L}_{cls} + \mathcal{L}_{mse} + \mathcal{L}_{im\text{-}cst} + \mathcal{L}_{vlprox}$	0.17 / 0.39	0.59 / 1.17	2.17 / 4.20

$$\mathcal{M}_{vlalign}(\mathcal{X}, k) = \frac{\sum_{i=1}^{|\mathcal{X}|} \#\mathrm{reverse\_pairs}(\mathrm{arrS}(i, k))}{|\mathcal{X}|}$$

$$\mathrm{arrS}(i, k) = [||S(\mathbf{x}_i) - T_{txt}(l(y_j))||_2]_{j \in \mathcal{I}(i,k)}$$

$$\mathcal{I}(i, k) = \mathrm{argtopk}([-||T_{im}(\mathbf{x}_i) - T_{txt}(l(y_j))||_2]_{j=1}^{|\mathcal{Y}|})$$

Language Representation Enrichment

- **Prompt ChatGPT to enrich label descriptions:** "Use a single sentence to describe the appearance and shape of {cls}. Only describe the shape and appearance."
- **ChatGPT-enriched descriptions are significantly more informative than e.g., auxiliary captions generated by OFA**, supplying comprehensive details to distinguish finegrained categories.

Original Description:

"A photo of an Acura Integra Type R 2001"

ChatGPT-Enriched:

"A photo of an Acura Integra Type R 2001. The 2001 Acura Integra Type R features a compact and sporty design with a sleek, aerodynamic body, sharp angles, and a distinctive rear spoiler."

Auxiliary Caption from OFA:

"A white car is parked in a field."

- ChatGPT-enriched language representation space confers more **independent & meaningful attributes to distinguish labels**, allowing **OOD text features to be more precisely aligned with image features.**

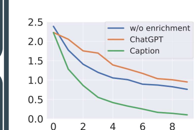
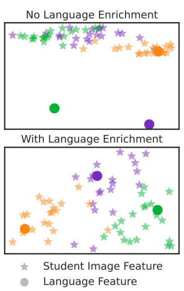


Figure 4: Top 10 eigenvalues of text features.

	Cos. Sim.
w/o enrichment	0.5156
ChatGPT	0.4462
Caption	0.5572

Table 6: Average cosine similarity between all pairs of text features.



Applications (e.g., Robotics)

	$\mathcal{X}_{id}$	$\mathcal{Y}_{id}$ on $\mathcal{X}_{ood}$	$\mathcal{Y}_{ood}$ on $\mathcal{X}_{ood}$
Closed-Set	96.4 / 96.5	NA / 86.2	NA / 87.7
$\mathcal{L}_{cls}$	96.9 / 97.2	79.3 / 85.3	71.7 / 87.3
+ $\mathcal{L}_{im\text{-}cst}$	**99.2 / 99.2**	84.0 / 91.9	76.3 / 88.3
+ Semantic Enrich	98.2 / 99.0	**84.3 / 92.0**	**83.0 / 89.6**

(a) Overall accuracy over all YCB objects

	$\mathcal{X}_{id}$	$\mathcal{Y}_{id}$ on $\mathcal{X}_{ood}$	$\mathcal{Y}_{ood}$ on $\mathcal{X}_{ood}$
Closed-Set	91.6 / 91.9	NA / 57.8	NA / 35.9
$\mathcal{L}_{cls}$	94.0 / 94.4	46.5 / 54.7	23.3 / 32.8
+ $\mathcal{L}_{im\text{-}cst}$	**98.1 / 98.0**	55.3 / **70.8**	**23.7** / 47.3
+ Semantic Enrich	97.2 / 97.4	**55.6 / 70.8**	11.7 / **50.7**

(b) F1-measure over objects that exist in observations.