Aligning Visual Contrastive Learning Models via Preference Optimization • ICLR 2025

Amirabbas Afzali, Borna Khodabandeh, Ali Rasekh,
Mahyar JafariNodeh, Sepehr Kazemi, Simon Gottschalk
L3S Research Center

Introduction & Motivation

- Vision-Language Models (e.g., CLIP) are powerful but vulnerable to some attacks like typographic attack and inherent biases.
- Aligning model behavior, in retrieval tasks, classification and downstream tasks with human preferences is crucial for fairness and robustness.
- Preference Optimization (PO) methods like RLHF, DPO, IPO, and KTO have been successful in generative models.

Background on Preference Optimization

- PO aims to train models to align with human preferences.
- Common methods include:
  - Reinforcement Learning from Human Feedback (RLHF): Uses a reward model trained on preferences to guide policy learning.
  - Direct Preference Optimization (DPO): Directly optimizes the policy based on preferences, without an explicit reward model.
  - Identity Preference Optimization (IPO): An alternative approach to directly optimizing the policy.
  - Kahneman-Tversky-Optimization (KTO): Another direct optimization method.
- These methods have shown success in aligning generative models.

Method in 30 Seconds

Core Idea

Teach CLIP to prefer human-aligned behaviors using AI alignment techniques

- **What's New**:
  - First application of PO methods to contrastive vision-language models
  - Simple training framework requiring only synthetic datasets with:
    - Problem cases (attacks/biases)
    - Normal (clean) examples
- **Key Feature**:
  - Adjustable ”concept knobs” after training
  - e.g., control gender bias strength

Problem Formulation

- We frame the problem as a Markov Decision Process (MDP).
- The learning task is modeled as: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \rho_0, R)$
- Components:
  - $s \triangleq x$
  - $a \triangleq y$
  - $\rho_0(s) \triangleq p(x)$
  - $R(s, a) \triangleq r(x, y)$
  - $\pi_\theta(a|s) \triangleq \frac{e^{f_\theta(y,x)}}{\sum_{y_i} e^{f_\theta(y_i,x)}}$
- Similarity score: $f_\theta(x, y) = \mathcal{I}_\theta(x)^T \mathcal{T}_\theta(y)/\tau$

Preference-Based Contrastive Optimization

- Policy ratio:

$$h_{\pi_\theta}(y_w, y_l, x) = (\log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x)) - (\log \pi_{\text{ref}}(y_w|x) - \log \pi_{\text{ref}}(y_l|x))$$

- Simplified for CLIP like models:

$$h_{\pi_\theta}(y_w, y_l, x) = \frac{1}{\tau}(\mathcal{I}_\theta(x) - \mathcal{I}_{\text{ref}}(x))^\top (\mathcal{T}(y_w) - \mathcal{T}(y_l))$$

- Preference objectives:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[-\log \sigma(\beta h_{\pi_\theta}(y_w, y_l, x))]$$

$$\mathcal{L}_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\left(h_{\pi_\theta}(y_w, y_l, x) - \frac{\beta^{-1}}{2}\right)^2\right]$$

Regularization

- We introduce a regularization term to ensure the trained model remains close to the reference model:

$$\mathcal{L}_{\text{reg}}(\pi, \pi_{\text{ref}}; \mathcal{D}_{\text{reg}}) = D_{\text{KL}}(\pi(y|x) \| \pi_{\text{ref}}(y|x)) = \mathbb{E}_{x \sim \mathcal{D}_{\text{reg}}} \mathbb{E}_{y \sim \pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right].$$

- The final loss function is defined as:

$$\mathcal{L}(\pi_\theta, \pi_{\text{ref}}; \mathcal{D}) = \mathcal{L}_{\text{pref}}(\pi_\theta, \pi_{\text{ref}}; \mathcal{D}_{\text{pref}}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\pi_\theta, \pi_{\text{ref}}; \mathcal{D}_{\text{reg}}).$$

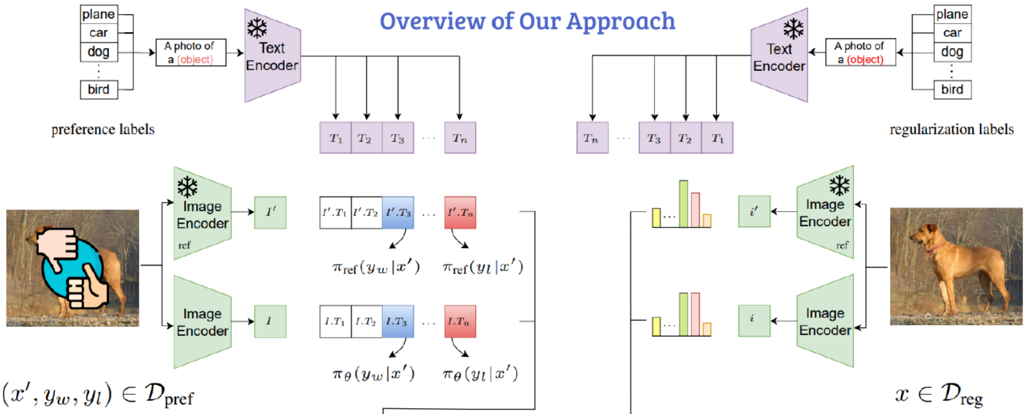
Linear Transformations and Adaptations

- Linear transformation matrix W to adjust the similarity function.
- SVD: $W = U\Sigma V^T$
- Modified similarity function:

$$\tilde{f}(y,x) = \mathcal{I}(x)^T W^T W T(y)/\tau = (V^T \mathcal{I}(x))^T \Sigma^2 (V^T T(y))/\tau$$

- Tune singular values using matrix powers: $W_t = U\Sigma^t V^T$

Overview of Our Approach



$$\mathcal{L}_{\text{pref}}(\pi_\theta, \pi_{\text{ref}}; \mathcal{D}_{\text{pref}}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\pi_\theta, \pi_{\text{ref}}; \mathcal{D}_{\text{reg}})$$

Experiments Results

- Experiments to evaluate effectiveness:
  - Typographic Robustness
  - Control between Optical Character Recognition (OCR) and Object Detection (OD)
  - Disentangling Gender Understanding

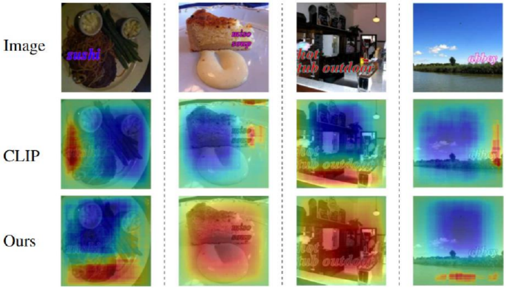
Typographic Robustness

Method	Caltech101		OxfordPets		StanfordCars		Flowers102		FGVCAircraft		DTD		SUN397		EuroSAT		Avg.	
	O	T	O	T	O	T	O	T	O	T	O	T	O	T	O	T	O	T
CLIP	88.64	63.97	87.35	58.95	58.72	21.02	66.32	31.32	18.99	10.83	44.57	25.53	61.74	34.02	42.98	4.86	58.66	31.31
Materzynska+	80.53	74.73	75.01	63.61	40.33	15.79	51.86	34.95	13.23	8.28	36.28	33.03	51.06	39.52	37.32	16.22	48.25	35.77
PAINT	88.48	83.57	85.23	76.53	55.30	33.44	**64.73**	54.92	17.73	14.46	**42.61**	36.60	61.69	53.62	38.20	17.31	56.74	46.31
Defense-Prefix	**89.28**	79.54	**87.22**	72.86	57.47	28.64	63.82	44.12	**19.26**	14.49	40.64	31.60	61.41	43.50	43.85	9.85	**57.87**	40.58
Ours (DPO)	87.50	85.43	85.25	79.72	56.03	34.33	56.60	55.70	16.21	13.87	39.36	38.48	61.02	56.34	**49.33**	28.32	56.41	49.02
Ours (IPO)	85.73	83.78	85.32	80.44	53.67	35.02	54.50	52.80	17.97	**15.86**	40.53	39.94	**61.91**	58.05	46.12	**43.23**	55.72	51.14
Ours (KTO)	87.67	**86.02**	85.41	**81.02**	**57.76**	**37.04**	59.10	**58.00**	17.27	15.59	40.74	**40.33**	**62.52**	**59.01**	46.26	36.94	57.09	**51.74**
Difference	↓1.61	↑2.45	↓1.81	↑4.47	↑0.9	↑3.60	↓5.63	↑3.08	↓1.99	↑1.10	↓1.87	↑3.73	↑0.83	↑5.39	↑2.41	↑19.63	↓0.78	↑5.43

Table 1: Classification accuracy on: O (Original dataset) and T (Typographic dataset).

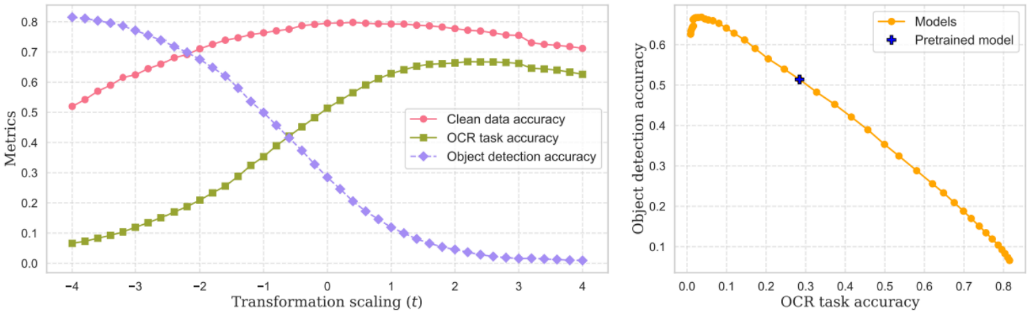
Typographic Robustness

(a) Retrieved images using VQGAN-CLIP using the captions *"focus"*, *"Love"*, *"Male police"* and *"Time"* for image generation.



(b) Saliency maps of vanilla CLIP and our fine-tuned model.

Control between OCR and OD

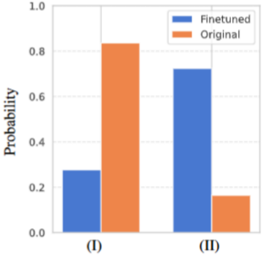


(a) Accuracy on typographic samples and percentage of typographic label predictions versus transformation scaling factor t. As t increases, the model favors object labels over typographic labels while maintaining accuracy.

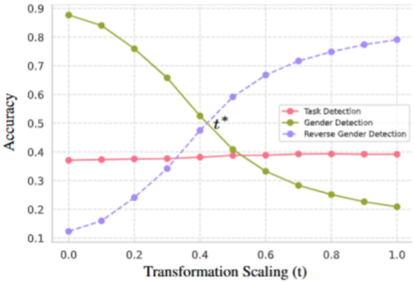
(b) Frontier of a DPO fine-tuned model, showing OCR vs. OD accuracy across with varying t.

Disentangling Gender Understanding

(a) Example image showing a man working.



(b) Model predictions before and after applying our gender-flipping method, showing changes in the predicted captions:
(I) “The man in the photo is working.“
(II) “The woman in the photo is working.“



(c) As t increases from 0 to 1, gender-specific predictions are reversed. t^* marks the point where gender information is neutralized, leading to balanced male and female predictions.

Disentangling Gender Understanding

Figure 4: Retrieved images for caption *"an image of a police"*, Top: Reversed(6W,4M), Middle: Original(2W, 8M), Bottom: Neutralized(5W, 5M) being the model at $t = t^*$

Conclusion

- We propose a novel approach to aligning and steering visual contrastive learning models with human preferences using Preference Optimization.
- Our method opens new avenues for developing more reliable and human-centered vision-language models.
- Additionally, this work provides insights into the principles and intuition behind Preference Optimization.

Contact

Amirabbas Afzali • L3S Research Center

amir8afzali@gmail.com

www.l3s.de