ΔEnergy: Optimizing Energy Change During Vision-Language Alignment Improves both OOD Detection and OOD Generalization

Lin Zhu [1] Yifeng Yang [1] Xinbing Wang [1] Qinying Gu [2] Nanyang Ye [1]

[1] Shanghai Jiao Tong University [2] Shanghai Artificial Intelligence Laboratory

Motivation

◆ In real-world scenarios, machine learning systems inevitably encounter both covariate shifts (e.g., changes in image styles) and semantic shifts (e.g., test-time unseen classes).

◆ Thus a critical but underexplored question arises: How to improve VLMs' generalization ability to closed-set OOD data, while effectively detecting open-set unseen classes during fine-tuning?

◆ Inspired by the substantial energy change observed in closed-set data when re-aligning vision-language modalities—specifically by directly reducing the maximum cosine similarity to a low value—we introduce a novel OOD score, named ΔEnergy.

◆ We show that ΔEnergy can simultaneously improve OOD generalization under covariate shifts, which is achieved by lower-bound maximization for ΔEnergy (termed EBM).

Problem Setting

Illustration of typical data setting in real-world scenarios.

Different Types of Data in Real World

Data	Closed-Set ID	Closed-Set OOD	Open-Set OOD
Image Input			
Classification	*Scarf*	*Scarf*	*Do Not Perform Classification*
Energy Changes	Huge ☺	Huge ☺	Small ☹

(i) closed-set ID data (e.g., dog);
(ii) closed-set OOD data with covariate shifts (e.g., dog with changed image styles);
(iii)open-set OOD data with semantic shifts (e.g., panda).

◆ The significant overlaps in energy distributions between closed-set ID and open-set OOD data pose a challenge for CLIP in detecting open-set OOD data.

◆ The notable discrepancy between the closed-set ID and closed-set OOD data also complicates achieving OOD generalization for closed-set OOD data.

Methodology

Motivated by the substantial energy change observed in closed-set data when re-aligning vision-language modalities, we propose a unified fine-tuning framework that optimizes this energy variation to enhance out-of-distribution generalization capability, while simultaneously enabling the model to detect unseen classes in open-set scenarios.

The proposed ΔEnergy, which measures the energy change after modifying the top-c maximum cosine similarities [3], unfolds as follows:

- Based on a pre-trained VLM, for each image feature $\mathbf{z_I}(\mathbf{x_i})$, we first select the text feature sets $\{\mathbf{h_j}(\mathbf{x_i})\}_{j=1}^{c}$ that have the top c similarity with $\mathbf{z_I}(\mathbf{x_i})$.
- We then compute the product between each image feature $\mathbf{z_I}(\mathbf{x_i})$ and the selected text feature $\mathbf{h_j}(\mathbf{x_i})$. The product feature is represented as $\mathbf{z_P}(\mathbf{x_i}, \hat{t}_j) = \mathbf{z_I}(\mathbf{x_i}) \odot \mathbf{h_j}(\mathbf{x_i})$.
- For each text feature $\mathbf{h_j}(\mathbf{x_i})$ ($j \in [1, \cdots, c]$), we denote the j-th largest cosine similarity between the image feature and the text feature as $s_{\hat{y}_j}(\mathbf{x_i}) = \mathbf{z_I}(\mathbf{x_i}) \cdot \mathbf{h_j}(\mathbf{x_i})$. Let $\tilde{s}_{\hat{y}_j}(\mathbf{x_i})$ represents the new cosine similarity after re-alignment, which is achieved by:

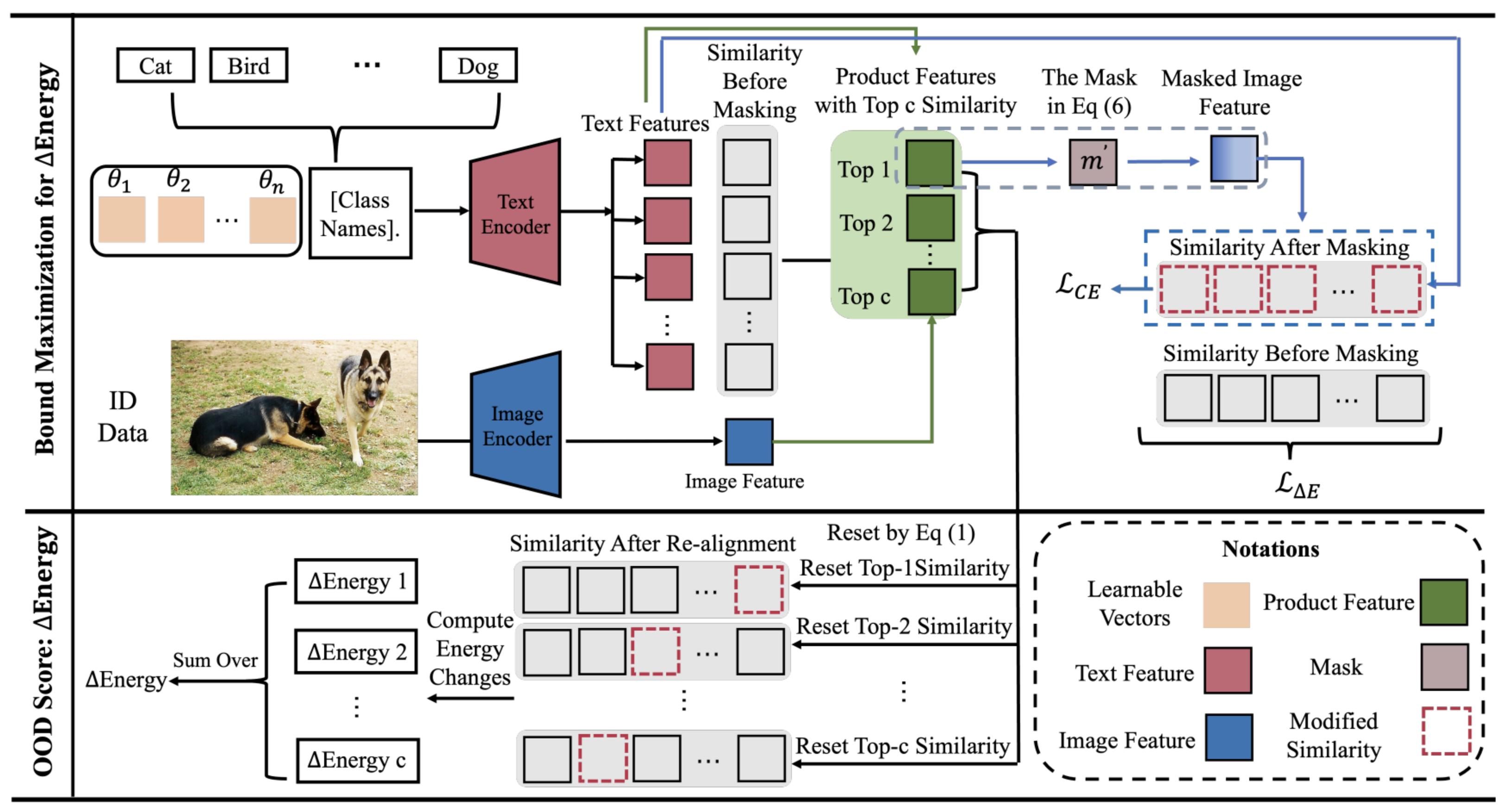
$$\tilde{s}_{\hat{y}_j}(\mathbf{x_i}) = 0 \tag{1}$$

- Finally, we can compute the new OOD score as: $\Delta\text{Energy}(\mathbf{x_i}) = E_1(\mathbf{x_i}) - E_0(\mathbf{x_i})$. Based on the scaling temperature τ, $E_0(\mathbf{x_i})$ is the energy score before the re-alignment:

$$E_0(\mathbf{x_i}) = -\log \sum_{j=1}^{K} e^{s_j(\mathbf{x_i})/\tau} \tag{2}$$

$E_1(\mathbf{x_i})$ is the energy score after the re-alignment:

$$E_1(\mathbf{x_i}) = -\frac{1}{c}\sum_{j=1}^{c} \log \left[e^{\tilde{s}_{\hat{y}_j}(\mathbf{x_i})/\tau} + \sum_{p \neq \hat{y}_j} e^{s_p(\mathbf{x_i})/\tau} \right] \tag{3}$$



Theorem 3.2. ***[OOD Detection Ability of ΔEnergy]*** *Suppose that the maximum cosine similarity for an ID sample $\mathbf{x}_{ID}$ is greater than that of an open-set OOD sample $\mathbf{x}_{OOD}$, i.e., $s_{\hat{y}_1}(\mathbf{x}_{ID}) > s_{\hat{y}_1}(\mathbf{x}_{OOD})$. Let $S_{\text{Method}}(\mathbf{x})$ denote the score assigned to sample $\mathbf{x}$ under a given method. We have the following properties: 1) $S_{\Delta\text{Energy}}(\mathbf{x}_{ID}) > S_{\Delta\text{Energy}}(\mathbf{x}_{OOD})$ for ID ($\mathbf{x}_{ID}$) and open-set OOD ($\mathbf{x}_{OOD}$) samples. 2) Compared to the MCM method, ΔEnergy amplifies the difference between ID and OOD data, i.e., $d_{\Delta\text{Energy}} > d_{MCM}$, where $d_{\text{Method}} = S_{\text{Method}}(\mathbf{x}_{ID}) - S_{\text{Method}}(\mathbf{x}_{OOD})$.*

Theorem 3.3. ***[The proposed OOD Score ΔEnergy gets lower FPR than MCM]*** *Given a task with closed-set ID label set $\mathcal{Y}_{\text{in}} = \{y_1, y_2, ..., y_K\}$ and a pre-trained VLM, for any test input $\mathbf{x}'$, based on the scaling temperature τ, the maximum concept matching (MCM) score is computed as follows:*

$$S_{\mathbf{MCM}}(\mathbf{x}'; \mathcal{Y}_{\text{in}}) = \max_i \frac{e^{s_i(\mathbf{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathbf{x}')/\tau}}.$$

For any $c \in \{1, 2, \cdots, K\}$, if $s_{\hat{y}_1}(\mathbf{x}') \leq \tau \ln 2$, we have

$$\text{FPR}^{\Delta\text{Energy}}(\tau, \lambda) \leq \text{FPR}^{\text{MCM}}(\tau, \lambda),$$

where $\text{FPR}^{\Delta\text{Energy}}(\tau, \lambda)$ and $\text{FPR}^{\text{MCM}}(\tau, \lambda)$ is the false positive rate of ΔEnergy and MCM, respectively, based on the temperature τ and detection threshold λ.

Theorem 3.5. ***[EBM leads to domain-consistent Hessians]*** *Given the ID training data sampled from domain $\mathcal{S}$ and the learnable parameter θ in VLM, we denote the masked domain as $\mathcal{S}'$. We represent the empirical classification loss on the domain $\mathcal{D}$ as $\widehat{\mathcal{E}}_{\mathcal{D}}(\theta)$. Let $\widehat{\mathbf{G}}_{\mathcal{D}}(\theta)$ and $\widehat{\mathbf{H}}_{\mathcal{D}}(\theta)$ be the gradient vector and Hessian matrix of empirical risk $\widehat{\mathcal{E}}_{\mathcal{D}}(\theta)$ over parameter θ, respectively. In this paper, we propose to minimize $\mathcal{L}_{\Delta E}$. The distance between the unmasked and masked image feature is assumed to satisfy: $\|\mathbf{z_I}(\mathbf{x_i}) - (\mathbf{z_I}(\mathbf{x_i}) \odot \mathbf{m}'(\mathbf{x_i}))\|_2 \leq \varepsilon$. Then the local optimum θ of $\min \mathcal{L}_{\Delta E}$ satisfies:*

$$|\theta^\top(\widehat{\mathbf{H}}_{\mathcal{S}}(\theta) - \widehat{\mathbf{H}}_{\mathcal{S}'}(\theta))\theta| \leq \frac{\varepsilon}{N}\sum_{i=1}^{N} |\theta^\top \nabla_\theta^2 \mathbf{z_T}(\mathbf{x_i})\theta|$$

Proposition 3.6. ***[EBM bound OOD generalization]*** *Let $\mathbf{z_I}(\mathbf{x_i})$ and $\tilde{\mathbf{z}}_\mathbf{I}(\mathbf{x_i})$ denote the image feature from source domain ($\mathcal{S}$) and target domain ($\mathcal{T}$), respectively. We assume that $||\mathbf{z_I}(\mathbf{x_i}) - \tilde{\mathbf{z}}_\mathbf{I}(\mathbf{x_i})||_2 \leq \varepsilon_1$. By applying the second-order Taylor expansion and utilizing the domain-consistent Hessians as outlined in Theorem 3.5, the OOD generalization gap between source domain ($\mathcal{S}$) and target domain ($\mathcal{T}$) is upper bounded by the following inequality:*

$$\max_{\{\theta: |\widehat{\mathcal{E}}_{\mathcal{S}}(\theta) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| \leq \epsilon\}} |\widehat{\mathcal{E}}_{\mathcal{T}}(\theta) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| \lesssim |\widehat{\mathcal{E}}_{\mathcal{T}}(\theta^*) - \widehat{\mathcal{E}}_{\mathcal{S}}(\theta^*)| + \max \frac{1}{2}|\theta^\top \widehat{\mathbf{H}}_{\mathcal{S}}(\theta^*)\theta| + O(\varepsilon_1)$$

where θ^* is a local minimum across all domains, i.e., $\nabla_\theta \widehat{\mathcal{E}}_{\mathcal{D}}(\theta^*) = \mathbf{0}$.

Results

Algorithm OOD Score	CoOp MCM	CoCoOp MCM	CLIP-Adapter MCM	Bayes-CAL MCM	DPLCLIP MCM	CRoFT MCM	LoCoOp GL	NegPrompt GL	GalLoP GL	EBM (Ours) ΔEnergy
ID ACC ↑	82.11	81.59	79.91	82.31	82.46	82.03	82.14	81.46	**84.51**	81.52 (0.4)
OOD ACC ↑	61.36	62.58	60.58	61.95	61.53	62.83	61.18	60.39	61.75	**63.28 (0.2)**
AUROC ↑	72.94	76.38	74.86	74.44	72.81	76.30	70.03	60.86	56.97	**81.90 (1.9)**
FPR95 ↓	73.15	70.30	70.92	72.34	73.07	69.78	74.33	86.66	91.17	**65.90 (1.7)**