Language-Driven Anchors for Zero-Shot Adversarial Robustness
Supplementary Material

Xiao Li[1] Wei Zhang[1] Yining Liu[2] Zhanhao Hu[1] Bo Zhang[1] Xiaolin Hu[1,3]
[1]Department of Computer Science and Technology, BNRist, Institute for Artificial Intelligence,
IDG/McGovern Institute for Brain Research, Tsinghua Laboratory of Brain and Intelligence,
Tsinghua University, Beijing, China
[2]Harbin Institute of Technology, Weihai, China
[3]Chinese Institute for Brain Research (CIBR), Beijing, China
{lixiao20, zhang-w19, huzhanha17}@mails.tsinghua.edu.cn
22s030184@stu.hit.edu.cn
{dcszb, xlhu}@mail.tsinghua.edu.cn

A. Experimental Details on CIFAR100

We used a PreActResNet18 model on Cifar100 to investigate the influence of CLIP anchors on anchor-based AT performance. We used the CLIP ViT-B/16 text encoder. The prompt text was set as "This is a photo of a { }". Three types of anchors were evaluated: the original CLIP anchors, the expanded CLIP anchors (see Sec. 3.4), and the MMC anchors [10] (the average CoS < 0). The optimization objectives were set to maximize the CoS between output feature $\mathbf{z}$ and different types of GT anchors: $\frac{\mathbf{z}^T\mathbf{a}_y}{||\mathbf{z}||_2}$.

We adversarially trained three models with these anchors by generating adversarial examples via PGD with maximum adversarial perturbation $\epsilon = 8/255$ in l_∞-norm bound, with iterative steps $T = 10$ and the step size $2/255$. We evaluated the robustness of the models against PGD with 20 iterative steps and step size $2/255$. The model was optimized for 200 epochs by SGD with an initial learning rate of 0.1, a momentum of 0.9, and a weight decay of $5e^{-4}$. The learning rate was reduced by 10 two times after the 100th and 150th epochs. Fig. 3 shows the learning curves of the three models on Cifar100 test set.

B. The Derivation from Rotated Anchor to Expanded Anchor

For the unit hyper-sphere, $r = 1$. According to Eq. (1), the coordinates of the rotated anchor $\tilde{\mathbf{a}}_i$ can be written as

$$\begin{aligned}
\tilde{a}_i^{(1)} &= \cos\tilde{\phi}_i^{(1)};\\
\tilde{a}_i^{(2)} &= \sin\tilde{\phi}_i^{(1)}\cos\tilde{\phi}_i^{(2)};\\
&...\\
\tilde{a}_i^{(n)} &= \sin\tilde{\phi}_i^{(1)}\sin\tilde{\phi}_i^{(2)}...\sin\tilde{\phi}_i^{(n-2)}\sin\tilde{\phi}_i^{(n-1)},
\end{aligned} \tag{1}$$

where $\tilde{a}_i^{(j)}$ denotes the j-th element of $\tilde{\mathbf{a}}_i$, and $\tilde{\phi}_i^{(j)}$ denotes the j-th angular coordinates anchor $\tilde{\mathbf{a}}_i$. Once fixed ϕ_0, we expand the polar angle as $\bar{\phi}_i^{(1)} = \frac{\pi}{2}\cdot\frac{\tilde{\phi}_i^{(1)}}{\phi_0}$, while angles with other coordinates remain unchanged, *i.e.*, $\bar{\phi}_i^{(j)} = \tilde{\phi}_i^{(j)}, j = 2, ..., n$. Then we can deduce the expressions for the coordinates of the expanded anchor as:

$$\begin{aligned}
\bar{a}_i^{(1)} &= \cos\bar{\phi}_i^{(1)};\\
\bar{a}_i^{(2)} &= \sin\bar{\phi}_i^{(1)}\cos\bar{\phi}_i^{(2)}\\
&= (\sin\tilde{\phi}_i^{(1)}\cos\tilde{\phi}_i^{(2)})\cdot\sin\bar{\phi}_i^{(1)}/\sin\tilde{\phi}_i^{(1)}\\
&= \tilde{a}_i^{(2)}\cdot\sin\bar{\phi}_i^{(1)}/\sin\tilde{\phi}_i^{(1)}\\
&...\\
\bar{a}_i^{(n)} &= \sin\bar{\phi}_i^{(1)}\sin\bar{\phi}_i^{(2)}...\sin\bar{\phi}_i^{(n-1)}\\
&= (\sin\tilde{\phi}_i^{(1)}\sin\tilde{\phi}_i^{(2)}...\sin\tilde{\phi}_i^{(n-1)})\cdot\sin\bar{\phi}_i^{(1)}/\sin\tilde{\phi}_i^{(1)}\\
&= \tilde{a}_i^{(n)}\cdot\sin\bar{\phi}_i^{(1)}/\sin\tilde{\phi}_i^{(1)}.
\end{aligned} \tag{2}$$

Thus, we the Cartesian coordinates of expanded anchors $\bar{\mathbf{a}}_i$ are given by:

$$\begin{aligned}
\bar{a}_i^{(1)} &= \cos\bar{\phi}_i^{(1)};\\
\bar{a}_i^{(j)} &= \tilde{a}_i^{(j)}\cdot\sin\bar{\phi}_i^{(1)}/\sin\tilde{\phi}_i^{(1)}, \quad j = 2, \cdots, n.
\end{aligned} \tag{3}$$

C. Pseudo Code of Expansion Algorithm

The pseudo-code of the expansion algorithm can be found in Algorithm 1. Given the original anchors $\{\mathbf{a}_i\}_{i=1}^N$, we can obtain the expanded final anchors $\{\hat{\mathbf{a}}_i\}_{i=1}^N$ by this algorithm.

Algorithm 1: Expansion Algorithm

```
Input  : The original anchors $\{\mathbf{a}_i\}_{i=1}^{N}$, where
         $\mathbf{a}_i \in \mathbb{R}^n$ and $N$ denotes the number of
         anchor points.
Output: The expanded text anchors $\{\hat{\mathbf{a}}_i\}_{i=1}^{N}$.
Find a center: $\mathbf{v} \leftarrow \frac{\sum_{i=1}^{N} \mathbf{a}_i}{||\sum_{i=1}^{N} \mathbf{a}_i||_2}$
Calculate a rotation matrix $\mathbf{R}$ so that $\mathbf{Rv} = \mathbf{p}$, where
  $\mathbf{p} = [1, 0, \cdots, 0]$
for $i = 1$ to $N$ do
    $\tilde{\mathbf{a}}_i \leftarrow \mathbf{Ra}_i$
end
$\phi_0 \leftarrow \max_{1 \le i \le N} \{\arccos \tilde{a}_i^{(1)}\}$, where $\tilde{a}_i^{(j)}$ denotes
  the $j$-th element of $\tilde{\mathbf{a}}_i$
for $i = 1$ to $N$ do
    $\tilde{\phi}_i^{(1)} \leftarrow \arccos \tilde{a}_i^{(1)}$
    $\bar{\phi}_i^{(1)} \leftarrow \frac{\pi}{2} \cdot \frac{\tilde{\phi}_i^{(1)}}{\phi_0}$
    $\bar{a}_i^{(1)} \leftarrow \cos \bar{\phi}_i^{(1)}$
    for $j = 2$ to $n$ do
        $\bar{a}_i^{(j)} \leftarrow \tilde{a}_i^{(j)} \cdot \sin \bar{\phi}_i^{(1)} / \sin \tilde{\phi}_i^{(1)}$
    end
end
for $i = 1$ to $N$ do
    $\hat{\mathbf{a}}_i = \mathbf{R}^{\mathbf{T}} \bar{\mathbf{a}}_i$
end
Return: $\{\hat{\mathbf{a}}_i\}_{i=1}^{N}$.
```

Hyper-Parameter	Clean	Robust
$\tau = 0.03$	44.37	32.51
$\tau = 0.07$	51.43	35.23
$\tau = 0.2$	58.80	37.52
$\tau = 0.5$	56.07	40.17
$\tau = 1.0$	55.60	40.12

Table S1. Classification accuracy (%) of Conv4-512 on CIFAR-FS with different τ. Here all experiments are performed based on expanded anchors with A-CE loss and smoothness loss supervision. The robust accuracy is evaluated by PGD-20.

D. The Influence of τ

As stated in Sec. 3.5, several previous works [8, 12] on standard training used a temperature parameter $\tau = 0.07$ to scale the CoS, *i.e.*, L_1 becomes:

$$L_1^{\star} = \mathbb{E}_{(\mathbf{x},y)} \left[-\log \frac{\exp(f_{\Theta}(\mathbf{x}+\delta)^T \hat{\mathbf{a}}_y / \tau)}{\sum_{i=1}^{N} \exp(f_{\Theta}(\mathbf{x}+\delta)^T \hat{\mathbf{a}}_i / \tau)} \right]. \quad (4)$$

We empirically observed that this might be harmful to the zero-shot performance under AT. Tab. S1 shows the classification accuracy of Conv4-512 on CIFAR-FS with different τ in the 5-way zero-shot setting (similar to those in Tab. 3). We can see that the model with smaller τ has worse adversarial robustness in the zero-shot setting. Thus, for simplicity, we set $\tau = 1$ in all other experiments.

E. Introduction to Downstream Datasets

We evaluate the zero-shot adversarial robustness trained on ImageNet-1K on ten downstream datasets, covering a diverse range of recognition tasks. AwA2 [13] and aPY [4] are two popular datasets in the ZSL setting. COCO Objects are images extracted from the bounding box annotations of MS COCO [9]. We also include Cifar100 [7], STL10 [3], Caltech101 [5], and Caltech256 [6] for generic classification; OxfordPet [11] for fine-grained classification; DTD [2] for texture recognition; and SUN [14] for scene recognition.

F. Smoothness v.s. TRADES

Smoothness in LAAT is different from TRADES [15] in several aspects. First, TRADES tries to minimize the classification loss on benign examples and the KL divergence between outputs of benign and adversarial examples, while we try to minimize the classification loss on adversarial examples and maximize the Cosine Similarities (CoS) between benign examples and adversarial examples. Second, the adversarial generation of TRADES is derived from the KL divergence, while the adversarial generation of LAAT only uses the classification loss (L_1).

G. Semantic Consistency v.s. Zero-Shot Performance

We used different CLIP text encoders to investigate the relationship between the semantic consistency of text encoders and the zero-shot adversarial performance. We performed experiments on Cifar100, which has 100 categories belonging to 20 super-categories. Each super-category includes 5 categories, *e.g.*, *apple*, *mushroom*, *sweet pepper*, *orange*, and *pear* belong to the same super-category *fruit and vegetables*. The categories within the same super-category are semantically similar categories. Therefore, a text encoder with high semantic consistency should map the categories within the same super-category to neighboring anchors and these anchors should have high CoS. We designed two metrics to measure the semantic consistency of a text encoder by using the super-categories.

We first calculated CoS between the category anchors obtained from a text encoder and then numbered the categories in descending order of the CoS. The numbers ranged between 0 to 99. We name them as *ranks* next. Tab. S2 shows the ranks of five categories in the super-category *fruit and vegetables* with text encoder ViT-B/16. For example, the *apple* anchor's CoS with the *pear* anchor is the 2nd highest among its 99 CoS with other anchors (except the

Category	Apple	Mushroom	Sweet pepper	Orange	Pear
Apple	0	7	1	27	2
Mushroom	7	0	4	23	34
Sweet pepper	1	11	0	32	10
Orange	4	10	3	0	2
Pear	2	30	3	6	0

Table S2. The ranks of categories belonging to the super-category *fruit and vegetables* with text encoder ViT-B/16. Each rank is calculated with the row category and all column categories. Each category has the highest CoS with itself, so the ranks on the diagonal are always 0.

Text Encoder	Sum of rank	Top-5 ratio
RN50x4	332	54.6%
RN50x16	319	54.4%
ViT-B/32	286	60.6%
ViT-B/16	**225**	**61.4%**
ViT-L/14	284	58.2%

Table S3. The average sum of ranks and the average top-5 ratio of ranks on 20 super-categories for each text encoder. The smaller sum of ranks and the larger top-5 ratio of ranks indicate greater semantic consistency of a text encoder.

apple anchor itself), so the rank in the first row and the fifth column of Tab. S2 is 2.

The two metrics were designed based on the ranks of the CoS described above. Intuitively, the higher the CoS within a super-category, the smaller the ranks within this super-category. Note that using ranks has an advantage over using the CoS directly, as ranks take the relationships between different super-categories into consideration. We calculated the sum of ranks and top-5 ratio of ranks within each super-category, and averaged them over all 20 super-categories as the two metrics. Take *fruit and vegetables* in Tab. S2 as an example, the former is the sum of $5 \times 5 = 25$ ranks, 219 in this case, and the latter is the ratio of ranks from 0 to 4 out of 25 ranks, $14/25 = 56\%$ in this case. The smaller sum of ranks and the larger top-5 ratio of ranks indicate greater semantic consistency of a text encoder.

Tab. S3 shows each text encoder's average sum of ranks and the average top-5 ratio of ranks on the 20 super-categories. We can see that ViT models are better than ResNet models under these two metrics. These results on semantic consistency generally correspond to the results on zero-shot adversarial robustness in Tab. 7 (note that CIFAR-FS [1] used in Tab. 7 is a variant of Cifar100), *e.g.*, ViT-B/16 has the best semantic consistency under these two metrics and it also has the best zero-shot adversarial robustness among these models. This correspondence indicates that the semantic consistency of the text encoder is one of the keys to zero-shot adversarial robustness.

References

[1] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Int. Conf. Learn. Represent. (ICLR)*, 2019. 3

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3606–3613, 2014. 2

[3] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011. 2

[4] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1778–1785, 2009. 2

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pages 178–178, 2004. 2

[6] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 2

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *California Institute of Technology*, 2009. 2

[8] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 2

[9] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 740–755, 2014. 2

[10] Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness. In *Int. Conf. Learn. Represent. (ICLR)*, 2020. 1

[11] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3498–3505, 2012. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn. (ICML)*, pages 8748–8763, 2021. 2

[13] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 41(9):2251–2265, 2019. 2

[14] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3485–3492, 2010. 2

[15] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Int. Conf. Mach. Learn. (ICML)*, pages 7472–7482, 2019. 2