Target Selection for Adversarial Evaluation of Vision Models Using Language Models

Katarzyna Filus
Institute of Theoretical and Applied Informatics,
Polish Academy of Sciences, Gliwice, Poland
Email: kfilus@iitis.pl

Jorge M. Cruz-Duarte
University of Lille, CNRS, Inria, Centrale Lille,
UMR 9189 CRIStAL, F-59000 Lille, France
Email: jorge.cruz-duarte@univ-lille.fr

***Abstract*—Targeted adversarial attacks on vision models require selecting target labels, which strongly affects attack success. However, target selection is often treated as a minor detail in adversarial evaluation. Existing strategies typically rely on randomness, model predictions, or static semantic resources, which can limit interpretability, reproducibility, and flexibility. We propose a semantics-guided framework that transfers knowledge from pretrained language and vision-language models to guide practical target selection. We estimate inter-class similarity from pretrained embeddings and select the most and least semantically related labels relative to the ground truth, defining best-case and worst-case adversarial scenarios. Experiments with three language and vision-language models, three vision models, and five attack methods show that our approach produces practical targets and outperforms static lexical databases in the worst-case setting. We also find that static, target-only testing provides a preliminary assessment of the compatibility of similarity sources before attack execution. Our results support the use of pretrained language models as a principled basis for adversarial target selection and benchmarking.**

***Index Terms*—Artificial Intelligence Security, Adversarial Attacks, Target Label Selection, Neural Network Testing, Computer Vision, Semantic Similarity.**

I. Introduction

OVER the past decade, adversarial attacks have exposed vulnerabilities in deep vision models [1]–[5], complementing concerns posed by traditional security threats [6]–[8]. As attack methods mature, the focus should shift from proposing new attacks to building systematic evaluation protocols for security audits. In targeted attacks, though the choice of target labels is a critical factor that can strongly affect success, it is often treated only as an implementation detail. As a result, many studies rely on random or probability-based targets, which ignore the semantic structure of labels [3], [9], [10].

The dominant label-choice strategies are difficult to interpret, which limits transparency in adversarial testing. For instance, least-likely targets can depend on image-specific artifacts such as spurious background patterns, so it is often unclear why a given class is selected or whether it is semantically relevant. Because these strategies depend on model predictions for each image, targets vary across instances, which complicates verification and makes comparisons across models less meaningful. This lack of interpretability and scalability makes image-dependent target selection poorly suited for systematic benchmarking.

To build security-relevant adversarial benchmarks, target selection should be standardized and grounded in an interpretable structure, such as semantics. Prior work explored similarity-driven target selection using class structure from network weights and static lexical resources [11]. Static resources such as WordNet are a reasonable baseline, but they have practical limitations at scale. They require manual label-to-sense mapping and often provide coarse similarity resolution for distant classes, which can produce ties and unstable rankings [12]. Pretrained language and vision-language models offer an alternative by learning continuous semantic relations directly from data. Recent studies also report alignment between visual and language representations, which motivates cross-modal transfer for target selection [13]–[17].

We propose a unified framework for semantics-guided target label selection that uses pretrained language and vision-language models to estimate class similarity. Given a ground-truth label, we select the most similar label as the best-case target and the least similar label as the worst-case target, thereby producing easy and hard test cases. Because targets are defined at the class level, we can precompute lookup tables that remain fixed across images, thereby improving reproducibility and supporting standardized evaluation [11].

We also use the Dissimilarity Metric (DM) [5], [11] to quantify how far post-attack predictions move from the ground truth in the target model's class space. In addition, we evaluate a static variant of DM computed from target labels only, which provides a lightweight indicator of compatibility between a similarity source and a target model before running attacks. Moreover, we evaluate three similarity sources, three ImageNet-trained vision models, and five targeted attacks on a standard benchmark dataset [18]. The results show that pretrained embeddings yield more challenging worst-case targets than WordNet-based similarity, while vision-language embeddings are competitive for best-case targets.

The main contributions of this work are: (i) a semantics-guided, model-agnostic target selection method for targeted adversarial evaluation using pretrained text and vision-language embeddings; (ii) a comparison between pretrained language models and WordNet for best-case and worst-case target selection; (iii) an analysis of attack severity for different

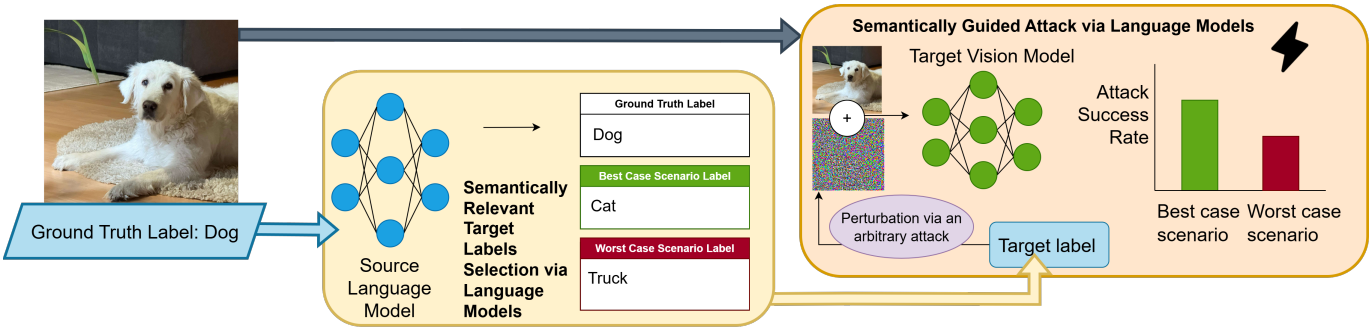


Fig. 1. Overview of the proposed semantics-guided target selection pipeline: a similarity source (text or vision-language) embeds class labels, ranks candidates relative to the ground-truth label, and returns MS and LS targets that define best- and worst-case evaluation scenarios for targeted attacks.

pretrained language models using DM, including a static variant that anticipates trends observed after attacks; (iv) Evidence that static similarity trends correlate with attack outcomes across target models.

As our proposal is more flexible than static language databases as a similarity source, our contribution constitutes a crucial step toward standardizing explainable security evaluation protocols. Lastly, the codes and results of this work are freely accessible at [19].

II. Related Work

There is a plethora of works on adversarial attacks in vision primarily focusing on crafting effective perturbations [1]–[3], [20]–[23], techniques to obtain transferability between models [24]–[26] and attack mitigation [27], [28]. In contrast, minimal attention has been paid to target label selection in targeted attacks, despite its central role in interpretability, reproducibility, and standardization in adversarial testing. The most common strategies include choosing random labels [3], [9], [10] and relying on probabilities of classes predicted for a given image [29]. In the literature, works implement predictions with the lowest- or highest-probability labels that are not the ground truth, i.e., Least Likely or Most Likely target labels [10], [29]. Alternatively, the K least- or most-likely labels can be used [10]. While these approaches are simple to implement, they inherently depend on instances. This renders the myriad resulting target labels uninterpretable and difficult to compare across models. Moreover, probability-based methods are sensitive to spurious background patterns and object co-occurrence, making it even more unclear why a target label was assigned to a specific image. Such a lack of transparency and scalability makes image-dependent strategies unsuitable for systematic benchmarking.

Moreover, proposals that connect security and explainability through semantics remain limited. Nevertheless, some works reveal the potential of semantics for adversarial testing. For example, Mahmood and Elhamifar proposed an optimization approach for multi-label learning that modifies the predictions of desired labels while ensuring that other labels are unaffected [30]. Other authors have considered semantics and other sources of similarity to build evaluation metrics [5], [31]. While these works show the potential of including similarity and semantics into adversarial testing, they focus on distinct aspects, such as metrics and multi-label optimization. In contrast, we focus on using similarity to efficiently and effectively guide adversarial testing.

A recent work introduced a method for target label selection that exploits similarities derived from internal model representations embedded in weights and lexical databases, such as WordNet, to define best- and worst-case testing scenarios [11]. In contrast to this work, we propose transferring knowledge from trained text- and text-image neural networks to guide cross-modal testing of vision models. Our motivation is that recent works in the literature reported the high alignment between different models and semantics [13]–[17], and that representations of knowledge largely overlap between models trained on different tasks and modalities [32]–[34]. In our approach, we use embeddings of known classes, generated from text-only label representations via text-only and text-vision models, to estimate concept similarity with low computational overhead and to test vision models. We exploit these similarities to build the best and worst adversarial cases by choosing the most and least similar target labels. This makes our procedure more flexible and more feasible to automate than using Wordnet similarities (work [11]), which requires manual label mapping. Moreover, there are many similarities and deficiencies across certain parts of these lexical databases. Also, using target-network weights introduces inconsistency in testing across different target models, thereby limiting reproducibility. Hence, our approach enhances the flexibility and consistency of testing. We use the WordNet-based approach presented in [11] as a reference in our experiments to assess the effectiveness of our method with respect to attack performance measures.

III. Methods

A. Semantic Similarity Reference

Semantic similarity measures how related two concepts are. We use WordNet [35] with the Wu-Palmer (WUP) measure [36], which is widely used for class-level semantic similarity [11], [31]. For a ground-truth class $c_{\text{gt}} \in \mathcal{C}$, we define c^{+}, the Most Similar (MS) or best-case, and c^{-}, the Least Similar (LS) or worst-case, as the classes with the highest and lowest WUP similarity to c_{gt}, respectively,

$$c^{\pm} = \operatorname{argext}_{c \in \mathcal{C} \setminus \{c_{\text{gt}}\}} \{\text{WUP}(c_{\text{gt}}, c)\}. \tag{1}$$

We compute the (c^+, c^-) pairs for all $c_{\text{gt}} \in \mathcal{C}$, where $\text{ext} \in \{\max, \min\}$, and store them in a lookup table.

B. Semantic Similarity via Pretrained Models

To estimate semantic similarity without WordNet, we encode each class name (e.g., "`school bus`") using pretrained text models (BERT, TinyLLAMA) or a vision-language model (CLIP). We refer to these as similarity source models. Let $\mathbf{e}(c) \in \mathbb{R}^d$ be the embedding of class $c \in \mathcal{C}$. We determine similarity between classes using cosine similarity $\text{cs}(\cdot)$. For a ground-truth class c_{gt}, the MS and LS targets are defined as,

$$c^{\pm} = \text{argext}_{c \in \mathcal{C} \setminus \{c_{\text{gt}}\}} \{\text{cs}(c_{\text{gt}}, c)\}. \quad (2)$$

We precompute and store the (c^+, c^-) pairs for all classes, with $\text{ext} \in \{\max, \min\}$, in a lookup table for each language model. This is performed once prior to evaluation and incurs negligible cost during testing, thereby facilitating large-scale security evaluations. The approach supports semantics-driven, consistent testing and improves interpretability and reproducibility.

C. Attack Evaluation Metrics

We evaluate targeted attacks using two standard metrics: Fooling Rate (FR) and Targeted Success Rate (TSR), computed on the attacked vision model. FR is the fraction of images whose predicted label changes, and TSR is the fraction for which the target label is reached. These metrics are defined as follows,

$$\text{FR} = \mathbb{E}\{\mathbb{1}(\hat{y}_i^{\text{pre}} \neq \hat{y}_i^{\text{post}})\} \text{ and } \text{TSR} = \mathbb{E}\{\mathbb{1}(\hat{y}_i^{\text{target}} = \hat{y}_i^{\text{post}})\}. \quad (3)$$

In both cases, $\hat{y}_i^k$, $\forall k \in \{\text{pre}, \text{post}, \text{target}\}$, stand for the i^{th} pre-attack, post-attack, and target labels, and $\mathbb{E}\{\cdot\}$ is expectation w.r.t. the uniform distribution over the N-sample dataset.

We also use the Dissimilarity Metric (DM), which measures attack severity in the model's class space [5]. DM quantifies the distance between the ground truth y_i and the post-attack prediction $\hat{y}_i$ based on class similarities induced by the target model's classifier weights. Let $y_i, \hat{y}_i \in \mathcal{Y}$ and let $r(y_i, \hat{y}_i)$ denote the corresponding rank distance. DM is defined as,

$$\text{DM} = \mathbb{E}\{d_i\} \quad \text{with} \quad d_i = r(y_i, \hat{y}_i)/(|\mathcal{Y}| - 1), \quad (4)$$

where $|\mathcal{Y}|$ is the number of classes and d_i is the normalized dissimilarity for sample i. DM lies in $[0, 1]$, where 0 indicates correct classification and 1 indicates the maximum rank distance in the model's class space. Compared with binary TSR, DM captures the severity of a misclassification even when the target label is not reached. This is relevant for LS (high DM) and MS (low, nonzero DM; about 0.001 is ideal for ImageNet). Using DM aligns with model-centric verification [11], [17], [37] and evaluations beyond label flipping [31], [38], [39].

We also compute DM in a static setting, using target labels instead of post-attack predictions. Here, $\hat{y}_i$ denotes the selected target for class i, and the mean is taken over all ImageNet classes ($N = |\mathcal{Y}|$). This static DM can be interpreted as a measure of compatibility between a similarity source and the target model. If static and post-attack DM show similar trends, vulnerability to a given target-selection source can be assessed before testing, without image inputs.

D. Experimental Setup

We evaluate five attacks: FGSM [2], C&W [3], PGD [21], Momentum Iterative Method (MIM) [40], and Simultaneous Perturbation Stochastic Approximation (SPSA) [41], using the same attack strengths as in [11]. We use three similarity source models: BERT [42], TinyLLAMA [43], and CLIP [44], from HuggingFace[1] implemented in PyTorch. As attack targets, we use MobileNetV2 [45], EfficientNetV2B0 [46], and ResNet50V2 [47], with accuracies of 83.3%, 89.7%, and 84.2%, respectively, on the test set. All models are ImageNet-trained [48], [49]. We test on the NIPS 2017 Adversarial Competition dataset [18], which is commonly used for adversarial evaluation [4], [50]. Moreover, the codes and results of this work are freely accessible at [19]. A portion of the experiments was carried out using a machine from the Grid'5000 testbed[2] with an NVIDIA Tesla P100 GPU with 16 GB RAM, and running on Linux. The other experiments were performed on a Windows-based workstation with an NVIDIA TITAN RTX GPU and 64 GB RAM.

IV. Experimental Results

This section reports results and answers for four research questions: (RQ1) reliability of text models for MS and LS target selection (Section IV-A); (RQ2) comparison with WordNet (Section IV-B); (RQ3) effect on attack severity via DM (Section IV-C); and (RQ4) static compatibility with target models (Section IV-D).

A. Reliability of Building Adversarial Cases with Text Models

Fig. 2 summarizes results for MS and LS targets across similarity sources. In the MS case (Fig. 2a), all sources yield higher FR and TSR than in the LS case (Fig. 2b). This is visible in the lower tails: MS remains above zero, whereas LS approaches zero for some runs. In MS, all sources achieve high FR (Avg. $\geq$ 0.7961). WUP and CLIP yield the highest average TSR (0.5563 and 0.5529), followed by BERT and LLAMA (0.4770 and 0.4592). CLIP and WUP also show lower variability, suggesting more stable target sets. In LS, FR averages range from 0.7745 (CLIP) to 0.7971 (WUP), and WUP and BERT exhibit the lowest variability (St. Dev. 0.3495 and 0.3759). TSR decreases across all sources (0.3919 to 0.4377), and CLIP and LLAMA yield the lowest average TSR, indicating more challenging worst-case targets.

TABLE I reports MS and LS results across similarity sources and target models. MS targets are consistently easier to reach, especially for weaker attacks, while the effect of target choice is smaller for stronger attacks. This suggests that language and vision-language models capture class relationships compatible with those of vision models, making them suitable for target selection and evaluation. TSR reflects target-selection effectiveness, and FR indicates that label semantics can also influence non-targeted success.

[1] https://huggingface.co/

[2] The Grid'5000 project is supported by a scientific interest group hosted by Inria and including CNRS, RENATER, several Universities, and other organizations. Further information: https://www.grid5000.fr.

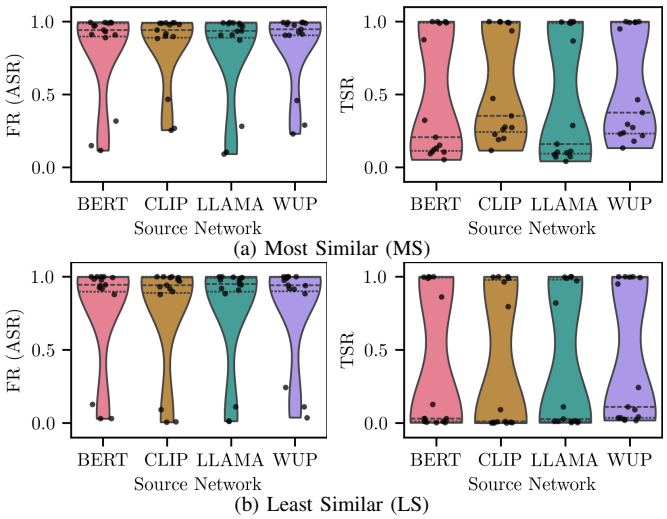


Fig. 2. Aggregated attack performance for (a) MS and (b) LS target selection strategies across all semantic similarity sources. Fooling Rate (FR) indicates overall label-change success, while Targeted Success Rate (TSR) reflects the success in reaching designated targets.

B. Comparing Similarity Sources for Target Label Selection

Using Fig. 2 and TABLE I, we identify when each similarity source is preferable. In LS, text and vision-language embeddings yield harder targets than WUP, as reflected by lower FR and TSR, consistent with expected better global distance resolution than WordNet. In MS, WUP achieves the highest average TSR (0.56), closely followed by CLIP (0.55), with similar average FR. BERT and TinyLLAMA yield high FR (0.80 and 0.81) but lower TSR, indicating slightly harder best-case targets. Overall, text-image models and WordNet perform slightly better on local similarity (MS), whereas text and vision-language models perform better on global similarity (LS). Thus, MS supports model-centric robustness evaluation, while LS is more informative for attack-centric evaluation.

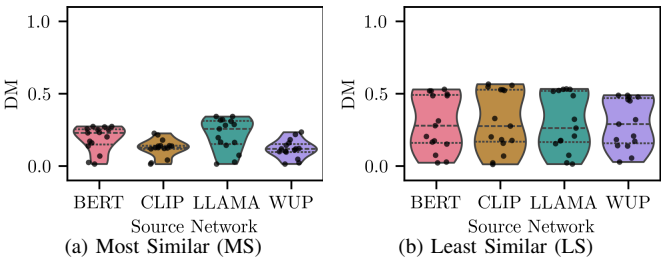


Fig. 3. Average Dissimilarity Metric (DM) for all similarity sources under (a) MS and (b) LS selection strategies. DM quantifies the rank-distance between the post-attack prediction and the ground-truth label in the attacked model's class space, capturing attack severity even when the target is not reached.

C. Impact of Similarity Sources on Attack Severity

Fig. 3 reports DM for each similarity source under MS and LS targets, capturing prediction drift even when TSR is low. As expected, DM is lower for MS and higher for LS. The results support the advantage of pretrained embeddings for LS targets: CLIP achieves the highest average DM (about 0.31), compared with about 0.30 for text-only models and 0.29 for WUP. Standard deviations are comparable across sources. For MS, CLIP remains close to WUP, whereas text-only models show larger dispersion. LS exhibits higher variability and wider ranges than MS. Maximum DM values are about 0.5 (0.6 for CLIP) for LS and about 0.2 to 0.3 for MS, indicating that LS can push predictions farther from the ground truth by roughly one-third of the model's rank distance. Therefore, we observe that CLIP and WordNet better capture local similarity, whereas text-only and vision-language models better support worst-case target selection.

TABLE I
FOOLING RATE (FR) AND TARGETED SUCCESS RATE (TSR) FOR MOST SIMILAR (MS) AND LEAST SIMILAR (LS) CASES ACROSS FIVE ATTACKS (CW, FGM, MIM, PGD, SPSA), THREE SOURCE MODELS, AND THREE TARGET MODELS. BEST VALUES PER ATTACK-VARIANT PAIR ARE IN BOLD. THE MODEL ABBREVIATIONS EFFICIENTN, MOBILEN, AND RESN CORRESPOND TO EFFICIENTNETV2, MOBILENETV2, AND RESNET50V2.

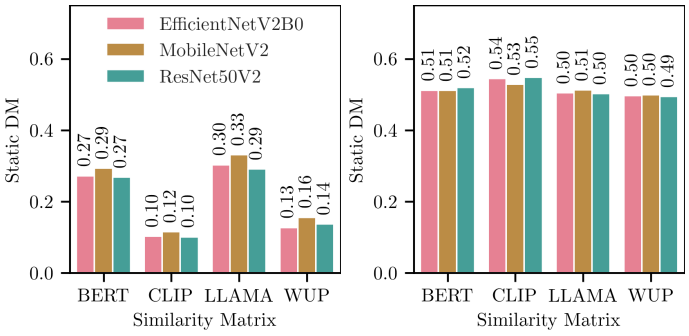
		EfficientN		MobileN		ResN		EfficientN		MobileN		ResN	
Att.	**Var.**	**FR**	**TSR**	**FR**	**TSR**	**FR**	**TSR**	**FR**	**TSR**	**FR**	**TSR**	**FR**	**TSR**
		BERT						CLIP					
CW	MS	0.15	0.15	0.32	0.32	0.12	0.12	0.27	0.28	**0.47**	**0.47**	**0.26**	**0.26**
	LS	0.03	0.03	0.13	0.13	0.03	0.03	**0.01**	**0.01**	**0.09**	**0.09**	**0.01**	**0.01**
FGM	MS	0.89	0.13	**0.93**	0.09	0.91	0.05	0.88	**0.27**	0.92	**0.19**	0.90	0.12
	LS	**0.88**	**0.01**	**0.94**	**0.00**	**0.92**	**0.00**	**0.88**	**0.01**	**0.94**	**0.00**	**0.92**	**0.00**
MIM	MS	0.99	0.99	0.99	**1.00**	0.97	0.88	0.99	0.99	0.99	**1.00**	**0.98**	0.94
	LS	1.00	1.00	1.00	1.00	0.98	0.86	1.00	0.97	1.00	1.00	0.98	0.80
PGD	MS	**1.00**	**1.00**	0.99	**1.00**	0.99	0.99	0.99	**1.00**	0.99	**1.00**	**0.99**	**0.99**
	LS	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**
SPSA	MS	0.91	0.10	**0.97**	0.21	0.94	0.11	0.90	0.20	**0.98**	0.36	0.94	**0.23**
	LS	0.93	0.01	0.98	0.03	0.95	**0.00**	**0.90**	**0.00**	**0.97**	**0.01**	**0.93**	**0.00**
		LLAMA						WUP					
CW	MS	0.10	0.11	0.28	0.29	0.09	0.10	**0.29**	**0.30**	0.46	0.46	0.23	0.24
	LS	**0.01**	**0.01**	0.11	0.11	**0.01**	**0.01**	0.04	0.04	0.24	0.24	0.11	0.11
FGM	MS	0.87	0.10	**0.93**	0.07	0.91	0.04	**0.91**	**0.27**	**0.93**	0.18	**0.91**	**0.13**
	LS	0.89	**0.01**	0.95	**0.00**	**0.92**	**0.00**	**0.88**	0.02	**0.94**	0.04	**0.92**	0.02
MIM	MS	0.99	0.99	0.99	**1.00**	0.98	0.87	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.95**
	LS	0.99	0.97	1.00	1.00	0.98	0.82	1.00	1.00	1.00	1.00	0.99	0.95
PGD	MS	0.99	**1.00**	0.99	**1.00**	0.99	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.99**
	LS	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**
SPSA	MS	0.91	0.07	**0.97**	0.16	0.94	0.09	**0.92**	**0.22**	**0.98**	0.38	**0.95**	**0.23**
	LS	0.91	0.01	**0.97**	0.03	0.95	0.01	0.92	**0.02**	0.98	0.09	0.94	0.04

D. Static compatibility of source and target models

Fig. 4 reports static DM between ground-truth and target labels, computed without images, and can be compared with post-attack DM in Fig. 3. TinyLLAMA yields the highest MS static DM, suggesting weaker local alignment, whereas CLIP and WUP yield the lowest MS static DM. These trends are consistent with FR/TSR in TABLE I, where CLIP performs best for MobileNetV2 and ResNet50V2. For LS, CLIP yields the highest static DM, supporting its ability to select worst-case targets. Thus, static DM provides an a priori compatibility check between similarity sources and vision models.

V. DISCUSSION

This work treats target label selection as a core design choice in adversarial evaluation, rather than an implementation detail,



(a) Most Similar (MS) (b) Least Similar (LS)

Fig. 4. Static Dissimilarity Metric (DM) computed for all ImageNet classes, based on MS and LS target labels from different semantic similarity sources.

and shows that semantics offers a practical path to standardization. By grounding targets in pretrained representations, the evaluation can be made more interpretable and reproducible, enabling class-level target sets that can be reused across images, attacks, and architectures.

The results consistently separate the best- and worst-case regimes: MS targets are reliably attainable across settings, whereas LS targets are substantially harder and lead to larger semantic drift. The source of similarity matters, but the overall trends are stable: CLIP is competitive with WordNet in the MS regime and renders stronger separation in the LS regime, suggesting that learned embeddings offer higher resolution for distant classes than lexical similarity. From a practitioner perspective, these observations support a simple recipe: use MS to probe model-centric robustness under semantically plausible confusions, use LS to stress-test attacks and characterize worst-case behavior, and use static DM as a low-cost screening step to select a compatible similarity source before launching full attack pipelines.

Several limitations remain. Because similarities are derived from class names, label ambiguity and naming conventions can influence rankings and may vary across datasets; moreover, the sources of similarity are not fully transparent, and DM relies on a weight-induced class geometry that may not reflect the full internal representation space. Future work will study more robust label handling (e.g., disambiguation and prompt variants), extend the approach to contextual and image-conditioned target selection (including open-vocabulary labels), and evaluate alternative severity measures that better capture geometry while preserving the efficiency required for large-scale benchmarking. We also plan to extend the framework to tasks beyond classification. We will also focus on extending the framework with techniques that could further improve the interpretability of testing, e.g., by incorporating recent advances from the mechanistic interpretability domain.

VI. Conclusions

We proposed a semantics-guided framework for targeted adversarial evaluation in which target labels are selected from pretrained text or vision-language embeddings rather than from image-dependent heuristics. For each ground-truth class, we estimate inter-class similarity in the embedding space and select the Most Similar (MS) label as the best-case (easy) target and the Least Similar (LS) label as the worst-case (hard) target. Because targets are defined at the class level, the resulting MS-LS pairs can be precomputed and stored in lookup tables, rendering fixed, interpretable target sets that remain consistent across images and architectures.

We validated the framework on a standard benchmark using three similarity sources (BERT, TinyLLAMA, CLIP), three ImageNet-trained target models, and five targeted attacks, and assessed performance using the Fooling Rate (FR), Targeted Success Rate (TSR), and the Dissimilarity Metric (DM). The experiments show that MS targets are consistently easier to reach than LS targets, supporting the use of MS for model-centric robustness checks and LS for attack stress-testing. We also found that CLIP is competitive with WordNet for best-case targets, while embedding-based sources yield more challenging worst-case targets and higher DM, indicating better resolution of global semantic distance than lexical similarity. Finally, the agreement between static DM (computed from targets only) and post-attack DM suggests a practical recommendation: use static DM as a lightweight, image-free screening step to select a compatible similarity source before running expensive attacks. Overall, the presented results demonstrate that our target selection approach provides a principled and effective basis for targeted adversarial evaluation.

Acknowledgments

This work was partially supported by the START Scholarship of the Foundation for Polish Science (FNP) for outstanding young scholars, agreement No START 017.2025, the Polish Minister of Science and Higher Education scholarship for outstanding young researchers, agreement No SMN/20/1546/2024. This work was partially supported by the National Centre for Research and Development (NCBR) and co-funded by the European Union under the European Funds for Modern Economy (FENG) Programme (SMART Konsorcja), project no. FENG.01.01-IP.01-A0GV/24-00, "Advanced maintenance and diagnostic tool for IT start-ups".

The authors utilized OpenAI's Prism, ChatGPT and Grammarly for language polishing under human supervision. The authors remain fully responsible for the manuscript.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *2nd International Conference on Learning Representations*, 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[4] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim, "Improving the transferability of targeted adversarial examples through object-based diverse input," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 244–15 253.

[5] K. Filus and J. Domańska, "Netsat: Network saturation adversarial attack," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 5038–5047.

[6] E. Gelenbe and M. Nasereddin, "Adaptive attack mitigation for iov flood attacks," *IEEE Internet of Things Journal*, 2025.

[7] M. Siavvas, I. Kalouptsoglou, E. Gelenbe, D. Kehagias, and D. Tzovaras, "Transforming the field of vulnerability prediction: Are large language models the key?" in *2024 32nd International Conference on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*. IEEE, 2024, pp. 1–6.

[8] K. Filus and J. Domańska, "Software vulnerabilities in TensorFlow-based deep learning applications," *Computers & Security*, vol. 124, p. 102948, 2023.

[9] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[10] S. Hu, L. Ke, X. Wang, and S. Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7649–7657.

[11] K. Filus and J. Domańska, "Similarity-driven adversarial testing of neural networks," *Knowledge-Based Systems*, vol. 305, p. 112621, 2024.

[12] K. Filus, M. Romaszewski, and M. Żarski, "Semantic depth matters: Explaining errors of deep vision networks through perceived class similarities," *arXiv preprint arXiv:2504.09956*, 2025.

[13] L. Ciernik, L. Linhardt, M. Morik, J. Dippel, S. Kornblith, and L. Muttenthaler, "Training objective drives the consistency of representational similarity across datasets," *arXiv preprint arXiv:2411.05561*, 2024.

[14] K. Filus and J. Domańska, "What is the doggest dog? examination of typicality perception in imagenet-trained networks," *Neural Networks*, vol. 188, p. 107425, 2025.

[15] Y. Gao, J. Liu, Z. Xu, J. Zhang, K. Li, R. Ji, and C. Shen, "Pyramidclip: Hierarchical feature alignment for vision-language model pretraining," *Advances in neural information processing systems*, vol. 35, pp. 35 959–35 970, 2022.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[17] K. Filus and J. Domańska, "How cnns and vits perceive similarities between categories," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2025.

[18] "Dataset for the NIPS 2017 adversarial competition," [online], -, available: https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset [Accessed: 2025-05-06].

[19] K. Filus and J. M. Cruz-Duarte, "Target Selection for Adversarial Evaluation of Vision Models Using Language Models - Codes, and Results," Jan. 2026. [Online]. Available: https://doi.org/10.5281/zenodo.18428159

[20] C. Li, H. Wang, W. Yao, and T. Jiang, "Adversarial attacks in computer vision: a survey," *Journal of Membrane Computing*, pp. 1–18, 2024.

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[22] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.

[23] Z. Chen, B. Li, S. Wu, K. Jiang, S. Ding, and W. Zhang, "Content-based unrestricted adversarial attack," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[24] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.

[25] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, "Transferable adversarial perturbations," in *European Conference on Computer Vision*, 2018, pp. 452–467.

[26] J. Wang, M. Wang, H. Wu, B. Ma, and X. Luo, "Improving transferability of adversarial attacks with gaussian gradient enhance momentum," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 421–432.

[27] K. Filus, J. Domańska, and J. Klamka, "Evaluating convolutional neural networks via measuring similarity of their first-layer filters and gabor filters," in *2024 32nd International Conference on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*. IEEE, 2024, pp. 1–8.

[28] R. Bayat and I. Rish, "Adversarial training with synthesized data: A path to robust and generalizable neural networks," in *ICML Next Generation of AI Safety Workshop*, 2024.

[29] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[30] H. Mahmood and E. Elhamifar, "Semantic-aware multi-label adversarial attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 251–24 262.

[31] K. R. Mopuri, V. Shaj, and R. V. Babu, "Adversarial fooling beyond" flipping the label"," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 778–779.

[32] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, "Zoom in: An introduction to circuits," *Distill*, vol. 5, no. 3, pp. e00 024–001, 2020.

[33] B. Chughtai, L. Chan, and N. Nanda, "A toy model of universality: Reverse engineering how networks learn group operations," in *International Conference on Machine Learning*. PMLR, 2023, pp. 6243–6267.

[34] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*. PMlR, 2019, pp. 3519–3529.

[35] P. Kolb, "Experiments on the difference between semantic similarity and relatedness," in *17th Nordic Conference of Computational Linguistics*, 2009, pp. 81–88.

[36] Z. Wu and M. Palmer, "Verb semantics and lexical selection," *arXiv preprint cmp-lg/9406033*, 1994.

[37] P. Biecek and W. Samek, "Model science: getting serious about verification, explanation and control of ai systems," *arXiv preprint arXiv:2508.20040*, 2025.

[38] J. Zhang, Z. Zhu, X. Wang, S. Liao, Z. Jin, F. Salim, and H. Chen, "Par-advgan: Improving adversarial attack capability with progressive auto-regression advgan," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2025.

[39] M. Shao, "Designing physical-world universal attacks on vision transformers," in *NeurIPS Safe Generative AI Workshop 2024*, 2024.

[40] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.

[41] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord, "Adversarial risk and the dangers of evaluating against weak attacks," in *International conference on machine learning*. PMLR, 2018, pp. 5025–5034.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[43] P. Zhang, G. Zeng, T. Wang, and W. Lu, "Tinyllama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[45] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[46] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, 2021, pp. 10 096–10 106.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[50] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, "Towards transferable targeted attack," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 641–649.