Qeios

Peer Review

Review of: "TagFog: Textual Anchor Guidance and Fake Outlier Generation for Visual Out-of-Distribution Detection"

Shiming Chen[1]

1. Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

The authors propose a novel method that leverages the rich semantic knowledge in large language models (LLMs) and large vision-language models (LVLMs) for out-of-distribution (OOD) detection. Extensive experimental results on multiple OOD detection datasets demonstrate the effectiveness of the proposed approach. This paper is well-organized, and the ideas are interesting. My concerns are listed below:

1. The authors propose FOG to generate fake OOD samples and demonstrate its effectiveness through ablation studies. However, the paper lacks comparative experiments with other methods for generating fake OOD samples. It would be beneficial if the authors could provide experimental results using other OOD sample generation approaches for a more comprehensive evaluation.

2. Since the authors take CLIP as the backbone, which is a zero-shot learning model, I encourage the authors to take a short discussion between zero-shot learning [a-d] and the OOD task in Section 1.

[a] Zero-Shot Learning: A Comprehensive Evaluation of the Good, the Bad and the Ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[b] Semantics-Conditioned Generative Zero-Shot Learning Via Feature Refinement. International Journal of Computer Vision, 2025.

[c] TransZero++: Cross Attribute-guided Transformer for Zero-Shot Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[d] EGANS: Evolutionary Generative Adversarial Network Search for Zero-Shot Learning. IEEE Transactions on Evolutionary Computation, 2024.

[e] GNDAN: Graph Navigated Dual Attention Network for Zero-Shot Learning. IEEE Transactions on Neural Networks and Learning Systems, 2024.

doi.org/10.32388/160GES

3. In TAG, the authors impose additional constraints on the visual encoder by aligning visual features with their anchors. However, this method appears to constrain only ID samples, ensuring that the visual encoder extracts features with richer semantic information. Given this, how does it enhance the model's OOD detection performance? The authors should provide more insights and explanations to clarify this point.

4. In the Model Inference section, could the authors provide a more detailed explanation of how the post-hoc OOD detection strategy assists the model's inference process? Additionally, the authors should present the default sample inference formula used in the model.

5. The authors claim that their method achieves SOTA performance. However, it seems to lack comparisons with recently proposed methods from the past year. If newer and better approaches have been introduced, the authors should include these works in Table 2 for a more comprehensive comparison.

6. The authors should provide potential limitations of the proposed method and possible future improvements to help guide further research in this area.

7. Many tables in the paper have formatting issues, and the authors should carefully check them, such as Table 2. Additionally, the layout of several tables makes them difficult to read and may cause confusion, for example, Tables 3 and 4.

8. The paper has poor formatting, with large blank spaces appearing on some pages. Additionally, some tables are placed too far from their corresponding textual descriptions, making it difficult to follow the content.

Declarations

Potential competing interests: No potential competing interests to declare.