

Improving Image Classification through Active Learning and Pseudo-Labeling in Vision-Language Models

Melhorando a Classificação de Imagens com Aprendizado Ativo e Pseudo-Rotulação em Modelos de Linguagem de Visual

Willian P. Amorim^{1*}, Cid Adinam Nogueira Santos², Priscila T. M. Saito³

Abstract: Visual Language Models (VLMs) combine natural language processing and computer vision to interpret multimodal data, such as images and text, showing great potential in image classification applications. This paper investigates the integration of Active Learning (AL) and pseudo-labeling techniques with VLMs to improve image classification in various domains. To achieve this, five AL strategies (Random Sampling, Uncertainty Sampling, Margin Sampling, Entropy Sampling, and Query-by-Committee) and three pseudo-labeling approaches (Direct, Confidence Threshold, and Feature Similarity) were evaluated iteratively. The results demonstrate that the combination of active learning and pseudo-labeling can achieve promising results, in addition to full class coverage in a few iterations. We conclude that the integration of AL with feature similarity-based pseudo-labeling offers a robust and efficient solution for image classification in limited-data scenarios, promoting high accuracy, class representativeness, and the reduction of propagation errors, with potential for applications in critical domains like healthcare and industry.

Keywords: visual language models — active learning — pseudo-labeling

Resumo: Os Modelos de Linguagem Visual (VLMs) combinam processamento de linguagem natural e visão computacional para interpretar dados multimodais, como imagens e texto, apresentando grande potencial em aplicações de classificação de imagens. Este artigo investiga a integração de técnicas de Aprendizado Ativo (AL) e pseudo-rotulação com VLMs para melhorar a classificação de imagens nas mais diversas áreas. Para isso, cinco estratégias de AL (Random Sampling, Uncertainty Sampling, Margin Sampling, Entropy Sampling e Query-by-Committee) e três abordagens de pseudo-rotulação (Direct, Confidence Threshold e Feature Similarity), foram avaliadas iterativamente. Os resultados demonstram que a combinação de aprendizado ativo e pseudo rotulação podem alcançar resultados promissores, além de cobertura total de classes em poucas iterações. Concluímos que a integração de AL com pseudo-rotulação baseada em similaridade de características oferece uma solução robusta e eficiente para classificação de imagens em cenários com dados limitados, promovendo alta acurácia, representatividade de classes e redução de erros de propagação, com potencial para aplicações em domínios críticos como saúde e indústria.

Palavras-Chave: modelos de linguagem visual — aprendizado ativo — pseudo-rotulação

¹Faculdade de Ciências Exatas e Tecnologia, Universidade Federal da Grande Dourados (UFGD), Brazil

³Departamento de Computação, Universidade Federal de São Carlos (UFSCar), Brazil

²Instituto de Pesquisas Eldorado, Brazil

*Corresponding author: willianamorim@ufgd.edu.br

DOI: <http://dx.doi.org/10.22456/2175-2745.150941> • Received: 01/09/2025 • Accepted: 26/09/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

The rapid advancement of Visual and Language Models (VLMs) has transformed tasks that require a joint understanding of visual and textual data, such as image captioning, visual question answering, and multimodal classification [1]. However, the success of these models heavily relies on large-scale and high-quality labeled datasets, a process that is often expensive

and time-consuming. In many real-world scenarios, only a limited number of labeled samples are available, while vast amounts of unlabeled data remain underutilized. This challenge is particularly prevalent in various domains that require specialized annotation, such as medical images or specific industrial applications [2].

From a technological standpoint, VLMs emerged due to limitations in computational vision and existing language

models [3, 4]. While traditional computer vision models are restricted to analyzing visual images and lack generative language capabilities, visual language models excel in handling language, text, and images, meaning VLMs offer benefits for both vision and language, making them even more versatile. On the other hand, Visual Language Models (VLMs) face complex challenges related to class imbalance, hallucination, and data scarcity, which directly impact their performance, accuracy, and reliability. Thus, developing methods that effectively leverage unlabeled data while minimizing the need for extensive manual labeling is a critical research problem in the field of multimodal machine learning.

Semi-supervised learning (SSL) and active learning (AL) have emerged as promising paradigms to address the scarcity of labeled data [5]. SSL techniques, such as pseudo-labeling, explore unlabeled data by assigning predicted labels (pseudo-labels), thereby expanding the training set [6]. However, the quality of pseudo-labels heavily depends on the performance of the initial classifier and the strategy used to assign them, which can introduce noise and degrade model performance if not carefully designed. Conversely, active learning focuses on selecting the most informative samples for human annotation, optimizing the labeling process by prioritizing samples that are likely to improve model performance [7]. By combining these approaches, it is possible to create an integrated framework that maximizes the utility of both labeled and unlabeled data, especially for VLMs, which benefit from robust multimodal representations.

Accordingly, we will propose an approach that integrates three pseudo-labeling strategies with five active learning techniques to enhance the training of VLMs in scenarios with limited labeled data, seeking to balance the trade-off between the quantity and quality of pseudo-labeled data. The pseudo-labeling strategies include: Direct, Confidence Threshold and Feature Similarity. For the active learning techniques, we will evaluate — Random Sampling, Uncertainty Sampling, Margin Sampling, Entropy Sampling, and Query-by-Committee. The use of pseudo-labeling will allow training to scale by incorporating unlabeled data, while active learning ensures that the limited labeling budget is spent on samples that maximize model improvement. By integrating these techniques, we aim to create a robust pipeline that leverages the best multimodal representations of VLMs to achieve high performance with limited labeled data.

This paper is organized as follows: Section 2 will introduce Vision Language Models, detailing their architecture and the inherent challenges in their application. In Section 3, we will detail the methodological proposal and the strategies to be developed, including the active learning and pseudo-labeling techniques used. Section 4 will present the experiments conducted and the analysis of the results obtained. Finally, the conclusion and directions for future work will be discussed in Section 5.

2. Vision Language Models

Vision Language Models integrate visual and textual data to perform tasks that involve understanding these two data modalities. These models combine natural language processing and computer vision techniques to interpret and process both images and text. They can be applied to various tasks, such as visual question answering, pseudo-labeling images, and generating descriptions for videos [8, 9].

Vision Language Models are classified into two groups: single-stream models and dual-stream models. This classification is based on how the different data modalities are combined. A single-stream Vision Language Model operates by combining visual and textual information into a single processing module. It adopts an early fusion of distinct modalities. In this process, feature vectors previously extracted from images and text are concatenated into a single combined vector. This unified representation is then fed into a single processing stream. One of the main advantages of this approach is its efficiency, as the same set of parameters is used for all modalities. This makes the model simpler and enhances computational efficiency during both training and inference. Examples of models that utilize the single-stream architecture include VisualBERT [10], Unicoder-VL [11], and UNITER [12].

In contrast, a dual-stream Vision Language Model processes visual and textual information in separate streams. These models use distinct encoders for each data modality: a pre-trained vision encoder, such as Convolutional Neural Networks or Visual Transformers for images, and a pre-trained text encoder based on Transformers for text. The representations are then fed into a multimodal fusion module, often using attention mechanisms, to create multimodal representations. This late fusion technique enables more complex interactions between modalities, capturing detailed dependencies between images and text. However, this architecture is more complex and computationally expensive compared to the single-stream approach. Examples of models that adopt the dual-stream approach include ViLBERT [13], LXMERT [14], and CLIP [1].

Encoder-only models are advantageous in scenarios where the primary goal is representation learning. They often exhibit streamlined processing and lower computational complexity, making them suitable for tasks requiring compact and informative representations. However, these models may lack the ability to generate detailed outputs, limiting their use in tasks that demand complex responses or creative generation. On the other hand, encoder-decoder models offer flexibility to generate diverse outputs, making them well-suited for tasks such as image captioning, translation, or any application requiring creative responses. The decoding step allows the transformation of joint representations into meaningful outputs. However, this versatility comes at the cost of increased computational load and complexity.

The combined modality representations may or may not pass through a decoder before producing the final output.

Therefore, a Vision Language Model can be either an encoder-only model or an encoder-decoder model. Encoder-only models are advantageous in scenarios where the primary goal is efficient representation learning, offering greater computational efficiency, lower complexity, and more compact representations. In contrast, encoder-decoder models provide the ability to generate more complex and diverse outputs, with greater detail and versatility, making them ideal for applications requiring content generation.

2.1 Challenges

In applications involving VLMs, the main challenges are class imbalance, hallucination and data scarcity. Class imbalance occurs when there is a disproportionate representation of examples from different categories within a dataset. In VLMs, this can particularly affect the model's ability to generalize fairly across various visually represented objects or concepts. Hallucination refers to the generation of incorrect or non-existent information, i.e., when the model creates or interprets data that is not present or supported by the input (such as images or text). When generating image descriptions or associating text with images, the model may "invent" details not present in the image, such as mentioning objects or scenes that do not exist. Regarding data scarcity, although many VLMs are trained on vast datasets, a lack of specific or diverse data can pose a problem, particularly in specialized domains or less common contexts. In scenarios where few images or texts are available for training, the model may fail to learn robust or accurate representations of certain categories, leading to poor performance in real-world applications.

To address these challenges, techniques such as pseudo-labeling and active learning can provide strategies aimed at improving model performance with less reliance on labeled data and greater control over the learning process. Pseudo-labeling is a technique within the context of semi-supervised learning where the model automatically labels unlabeled data based on its own predictions, and these labels are then used to retrain the model. Active learning is an approach where the model intelligently selects which data to label based on its uncertainty or which examples will be most useful for improving its performance. The model can interact with an oracle (i.e., a human or external source) to label only the most informative examples. The combination of pseudo-labeling, and active learning offers powerful approaches to address the challenges of class imbalance, hallucination, and data scarcity in Visual Language Models. These techniques enable more efficient and robust learning, reducing the need for large volumes of labeled data and helping to mitigate common issues in multimodal AI models.

3. Active Learning and Pseudo-Labeling Strategies

Active learning (AL) and pseudo-labeling are powerful techniques designed to enhance the performance of multi-class classification models when labeled data is scarce. In active

learning, the objective is to select the most informative samples from an unlabeled dataset $\mathcal{D}_u$ for annotation by an oracle, thereby constructing an annotated dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^L$, where $x_i \in \mathcal{X}$ represents an input sample (e.g., an image), $y_i \in \{1, \dots, K\}$ is the corresponding label, L is the number of labeled samples, and K is the number of classes. Pseudo-labeling complements this process by assigning predicted labels to samples in $\mathcal{D}_u$, expanding the training set without requiring additional oracle queries. This section presents five active learning strategies and three pseudo-labeling strategies used in our experiments, detailing their methodologies, theoretical foundations, and selection criteria.

3.1 Active Learning Strategies

Active learning strategies aim to select the most informative N samples from $\mathcal{D}_u$ to be labeled by the oracle, optimizing model performance with minimal annotation effort. The following strategies are evaluated:

Random Sampling. This baseline strategy randomly selects N samples from $\mathcal{D}_u$ with uniform probability. Formally, a subset of indices $\mathcal{S} \subseteq \{1, \dots, |\mathcal{D}_u|\}$ of size N is chosen without considering model predictions. This approach serves as a reference, as it does not exploit model uncertainty or other criteria, relying solely on randomness.

Uncertainty Sampling. Uncertainty sampling prioritizes samples where the model is least confident in its predictions. For a sample $x \in \mathcal{D}_u$, the model outputs a probability distribution $P(y|x; \theta)$ over the K classes, where θ denotes the model parameters. The uncertainty is measured as the maximum softmax probability, $\max_{y \in \{1, \dots, K\}} P(y|x; \theta)$, and the N samples with the lowest maximum probabilities (indicating highest uncertainty) are selected for labeling.

Margin Sampling. Margin sampling selects samples based on the difference between the probabilities of the two most likely classes. For a sample $x \in \mathcal{D}_u$, let $P(y_1|x; \theta)$ and $P(y_2|x; \theta)$ represent the probabilities of the top two predicted classes, where y_1 and y_2 are the indices of the highest and second-highest probabilities, respectively. The margin is defined as $P(y_1|x; \theta) - P(y_2|x; \theta)$, and the N samples with the smallest margins are chosen, as they reflect the most ambiguous predictions.

Entropy Sampling. Entropy sampling targets samples with the highest predictive entropy, capturing the overall uncertainty in the model's output distribution. For a sample $x \in \mathcal{D}_u$, the entropy is computed as:

$$H(x) = - \sum_{y=1}^K P(y|x; \theta) \log P(y|x; \theta).$$

The N samples with the highest $H(x)$ values are selected, prioritizing those with the most dispersed probability distributions.

Query-by-Committee (QBC). QBC employs an ensemble of M models, each with parameters θ_m ($m = 1, \dots, M$), to identify samples with high disagreement among the models. For a sample $x \in \mathcal{D}_u$, each model predicts a class $y_m \in$

$\{1, \dots, K\}$. The vote distribution is calculated as $V(y|x) = \frac{1}{M} \sum_{m=1}^M \mathbb{I}(y_m = y)$, where $\mathbb{I}$ is the indicator function. The vote entropy is:

$$H_v(x) = - \sum_{y=1}^K V(y|x) \log V(y|x).$$

The N samples with the highest $H_v(x)$ are selected, indicating significant disagreement among the committee.

3.2 Pseudo-Labeling Strategies

Pseudo-labeling assigns predicted labels to samples in $\mathcal{D}_u$, creating a pseudo-labeled dataset $\mathcal{D}_p = \{(x, \hat{y})\}$ to augment $\mathcal{D}_l$. The following three strategies are investigated, each balancing the trade-off between the quantity and quality of pseudo-labels:

Direct Pseudo-Labeling. This strategy assigns pseudo-labels to all samples in $\mathcal{D}_u$ using the direct predictions of the classifier trained on $\mathcal{D}_l$. For a sample $x \in \mathcal{D}_u$, the pseudo-label is $\hat{y} = \arg \max_{y \in \{1, \dots, K\}} P(y|x; \theta)$, and all samples are included in $\mathcal{D}_p$ without additional filtering. This approach assumes the classifier's predictions are sufficiently reliable, but it may introduce noise if the model is inaccurate, particularly when $\mathcal{D}_l$ is small.

Confidence Threshold (CT). The confidence threshold strategy assigns pseudo-labels only to samples in $\mathcal{D}_u$ where the maximum softmax probability exceeds a predefined threshold τ (e.g., $\tau = 0.8$). For a sample $x \in \mathcal{D}_u$, the pseudo-label is $\hat{y} = \arg \max_{y \in \{1, \dots, K\}} P(y|x; \theta)$, and the sample is included in $\mathcal{D}_p$ if $\max_y P(y|x; \theta) > \tau$. This method prioritizes the quality of pseudo-labels by selecting only high-confidence predictions, reducing noise but potentially limiting the number of samples in $\mathcal{D}_p$.

Feature Similarity. This strategy assigns pseudo-labels based on the similarity between feature representations of unlabeled and labeled samples, using a pre-trained model such as CLIP. For a sample $x \in \mathcal{D}_u$, its feature vector $f(x)$ is extracted (e.g., from CLIP's image encoder). The pseudo-label $\hat{y}$ is assigned as the label y_j of the labeled sample $(x_j, y_j) \in \mathcal{D}_l$ whose feature vector $f(x_j)$ is most similar to $f(x)$, typically measured by cosine similarity:

$$\text{sim}(x, x_j) = \frac{f(x) \cdot f(x_j)}{\|f(x)\| \|f(x_j)\|}.$$

The sample x is included in $\mathcal{D}_p$ if the similarity exceeds a threshold or based on the nearest neighbor. This approach leverages robust feature representations, reducing dependency on the classifier's initial accuracy and improving pseudo-label consistency by exploiting feature-level proximity.

These active learning and pseudo-labeling strategies are evaluated in our experiments to assess their impact on model performance under limited labeled data scenarios, as detailed in the subsequent sections. The combination of strategic sample selection via active learning and robust pseudo-labeling aims to maximize the utility of both labeled and unlabeled

data, enhancing classification performance in resource-constrained settings.

4. Experiment and Results

For the experiments and subsequent image classification tasks, we selected five publicly available image classification datasets: Mnist [15], Cifar-10 [16], Cifar-100 [16], FashionMNIST [17], and EuroSAT [18]. These datasets cover a variety of classification tasks involving different image patterns. In this work, the categorical label associated with the image is treated as the essential minimal textual element. Although it is a simplified form of text, it is the target of the classification and is crucial in how VLMs are utilized in classification tasks. To evaluate the three pseudo-labeling strategies, it's important to highlight the selection of five active learning approaches: Random Sampling, Uncertainty Sampling, Margin Sampling, Entropy Sampling, and Query-by-Committee (QBC). The datasets were fixedly split into 70% for training (including the unlabeled data pool and the initial labeled pool) and 30% for testing, while maintaining the same class distribution. In the training process, the active learning strategies were executed by selecting 10% to 50% of the labeled samples, and the same proportion of the unlabeled samples for pseudo-labeling. To serve as the basis for our image encoder, we adopted the ViT-B/32 and the CLIP model. Throughout all training steps, we used the SGD optimizer with a learning rate of 0.002, which is decreased by a cosine annealing scheduler. We also set the maximum number of epochs to 200. All methods were implemented with PyTorch 2.0.1 and executed on a single NVIDIA TITAN V.

4.1 Metrics

To validate the effectiveness of the proposed method, we will employ a comprehensive set of evaluation metrics. The performance will be evaluated by the mean of Accuracy, Selection Time of active learning and propagation error. The selection time refers to the time taken by the active learning technique to perform the sample selection process, while the propagation error relates to the pseudo-labeling error of the samples chosen by active learning. Computational efficiency will be analyzed by the total time per iteration ($T_{total} = T_{train} + T_{classification}$). The percentage of corrected samples per iteration ($P_{corrected}^{(i)}$) in iteration $i > 0$, (with i corresponding to 10%, 20%, 30%, 40%, and 50%) is given by $P_{corrected}^{(i)} = (N_{corrected}^{(i)} / N_{annotated}^{(i)}) \times 100\%$, where $N_{corrected}^{(i)}$ is the number of samples with labels corrected by the expert in iteration i , and $N_{annotated}^{(i)}$ is the total number of samples annotated in iteration i . In the first iteration ($i = 10\%$), $P_{corrected}^{(0)} = 100\%$. The coverage of the class space will be verified by the percentage of classes present per iteration $P_{classes}^{(i)} = C_{observed}^{(i)} / C_{total} \times 100\%$, where $C_{observed}^{(i)}$ is the number of distinct classes observed up to iteration i , and C_{total} is the total number of classes in the dataset. To illustrate the potential for performance improvement, we also measured the

performance when applying model learning with the entire dataset.

4.2 Results

We will be performing the analysis separated by pseudo-labeling strategy. The objective was to investigate the effectiveness of the Active Learning paradigms in optimizing the acquisition of labeled data to train a classification model. Figures 1, 4 and 7 show the accuracy obtained throughout the iterations. Notably, we included a reference line, “All data”, which represents the classification performance achieved with the full dataset. Figures 2, 5 and 8, illustrate the percentage of corrected samples per iteration ($P_{corrected}^{(i)}$), and Figures 3, 6 and 9, present the normalized average of Pseudo error, Selection Time and Time full Metrics for each dataset.

Analyzing the evolution of accuracy across each iteration reveals distinct patterns among all strategies. On MNIST, the Feature Similarity pseudo-labeling strategy consistently outperformed the others, with maximum accuracies of up to 96.24% (Entropy Sampling, 40% of selected labeled and pseudo-labeled data) and 96.16% (Margin Sampling, 50%), approaching the reference performance with the entire dataset (96.28%). For Cifar-10, Feature Similarity also showed superior results, with accuracies of up to 91.73% (Uncertainty Sampling, 50%), close to the 91.30% limit, highlighting the robustness of pre-trained representations like CLIP on more complex datasets. On Cifar-100, with a larger number of classes, Margin Sampling combined with Feature Similarity reached 63.90% (50%), surpassing other combinations but still distant from the 61.30% limit, suggesting challenges in scenarios with high class granularity. On FashionMNIST, Entropy Sampling with Feature Similarity obtained 85.26% (40%), exceeding the 84.30% limit, indicating efficiency in capturing discriminative representations. For EuroSAT, Feature Similarity with Margin Sampling achieved 87.26% (50%), practically equaling the 87.30% limit. The Confidence Threshold strategy generally outperformed Direct, but lagged behind Feature Similarity, which benefited from similarity in robust feature spaces. Among the Active Learning strategies, Margin Sampling and Query-by-Committee frequently showed superior performance, especially on more complex datasets.

The analysis of the percentage of classes present per iteration metric ($P_{classes}^{(i)}$) reveals notable situations in the ability of active learning strategies to cover the class space in the datasets. On Mnist, the Feature Similarity strategy combined with Uncertainty Sampling and Query-by-Committee achieved 100% coverage in the first iteration, while Random Sampling was consistently less effective, reaching only 60% in the first iteration with Direct. On Cifar-10, Margin Sampling with Feature Similarity achieved full coverage from the third iteration, outperforming Random Sampling, which reached 100% only in the fourth iteration with Direct. For Cifar-100, with 100 classes, Margin Sampling and Query-by-Committee with Feature Similarity stood out, reaching 95% and 92% in the first iteration, respectively, against 56% for Random Sampling with Direct, highlighting the superior

ity of uncertainty-based and feature-similarity strategies in complex datasets. On FashionMNIST, Margin Sampling and Uncertainty Sampling with Feature Similarity reached 100% in the first iteration, while Random Sampling with Direct was slower, reaching 70%. On EuroSAT, Margin Sampling, Entropy Sampling, and Query-by-Committee with Feature Similarity covered all classes from the first iteration, outperforming Random Sampling, which reached 80%.

The analysis of Selection Time, Pseudo Error, and Total Time reveals a clear trade-off between computational efficiency and pseudo-label quality. Random Sampling is the most computationally efficient strategy, with the lowest Selection Time and Total Time. However, it consistently results in the highest Pseudo Error, showing that its speed comes at the cost of less reliable pseudo-labels. In contrast, Margin Sampling combined with Feature Similarity consistently minimizes the Pseudo Error. This approach proves to be the most robust, offering a strong balance between minimizing propagation errors and maintaining acceptable computational efficiency, particularly on complex datasets. The Feature Similarity pseudo-labeling strategy, especially when combined with Margin Sampling, consistently stood out, achieving accuracies close to or better than the reference performance with all data (e.g., 96.24% on MNIST, 91.73% on Cifar-10, 87.26% on EuroSAT) and full class coverage in a few iterations (e.g., 100% in the first iteration on EuroSAT with Margin Sampling).

Furthermore, Margin Sampling with Feature Similarity minimized the Pseudo Error and maintained a competitive Time Full, demonstrating its robustness on complex datasets like Cifar-100, where it achieved 63.90% accuracy and 95% class coverage in the first iteration. In contrast, Random Sampling, although computationally efficient with the lowest Selection Time and Time Full, showed a higher Pseudo Error and lower class coverage, making it less effective in scenarios requiring representativeness and precision. Query-by-Committee also showed promising results, but its higher computational complexity limits its practical applicability. Thus, the combination of Margin Sampling with Feature Similarity emerges as the most effective approach to maximize accuracy, class coverage, and pseudo-label quality, while maintaining reasonable computational efficiency, being particularly recommended for datasets with high class granularity and labeled data constraints. These findings reinforce the importance of uncertainty-based strategies and robust feature representations in semi-supervised learning, offering a promising path for real-world applications with limited resources.

5. Conclusion

In our study, we conducted a comprehensive evaluation with different optimization scenarios to train classification models based on Visual Language Models (VLMs), exploring the potential of active learning and pseudo-labeling, with a focus on maximizing performance in limited-data scenarios. For this, we investigated five AL strategies — Random

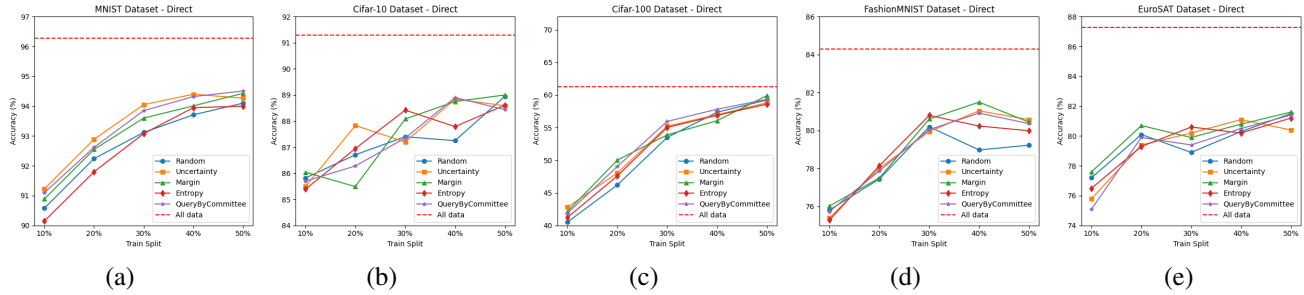


Figure 1. Direct Pseudo - Strategy : Accuracy per round comparison of Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

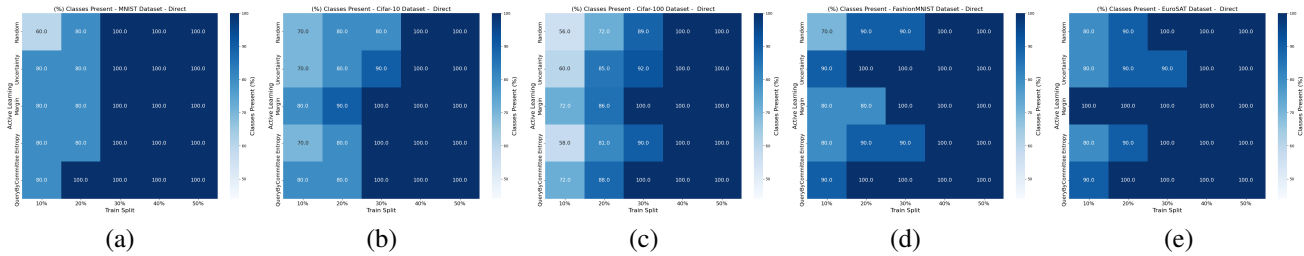


Figure 2. Direct Pseudo - Strategy : Percentage of classes present ($P_{classes}^i$) per iteration of Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

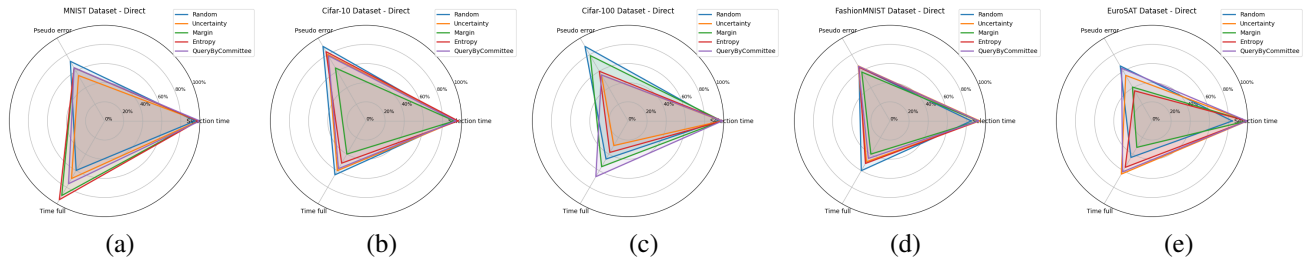


Figure 3. Direct Pseudo - Strategy : Pseudo error, Selection Time and Time full Metrics for each dataset, presented in radar charts Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

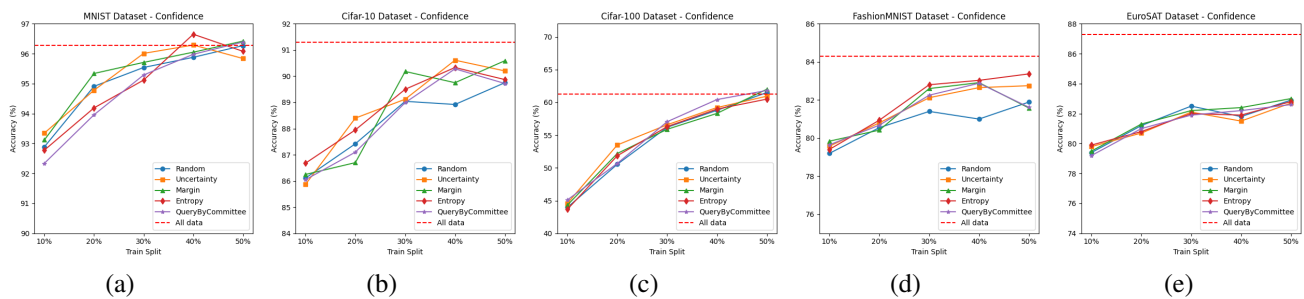


Figure 4. Confidence Pseudo - Strategy : Accuracy per round comparison of Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

Sampling, Uncertainty Sampling, Margin Sampling, Entropy Sampling, and Query-by-Committee — and three pseudo-labeling approaches — Direct, Confidence Threshold, and Feature Similarity — applied to the Mnist, Cifar-10, Cifar-100, FashionMNIST, and EuroSAT datasets.

These strategies were analyzed with the aim of exploring the strategic selection of representative subsets of labeled

data and the expansion of the training set through pseudo-labels. The results obtained from our experiments consistently demonstrate the effectiveness of the proposed approaches, particularly the superiority of those that combine Margin Sampling with Feature Similarity. We observed that the strategic selection of a small, representative dataset through active learning, combined with the expansion of the training set via

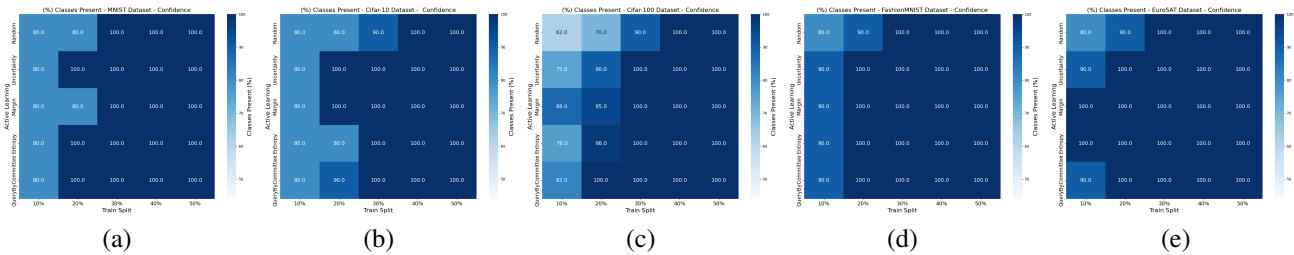


Figure 5. Confidence Pseudo - Strategy : Percentage of classes present ($P_{classes}^i$) per iteration of Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

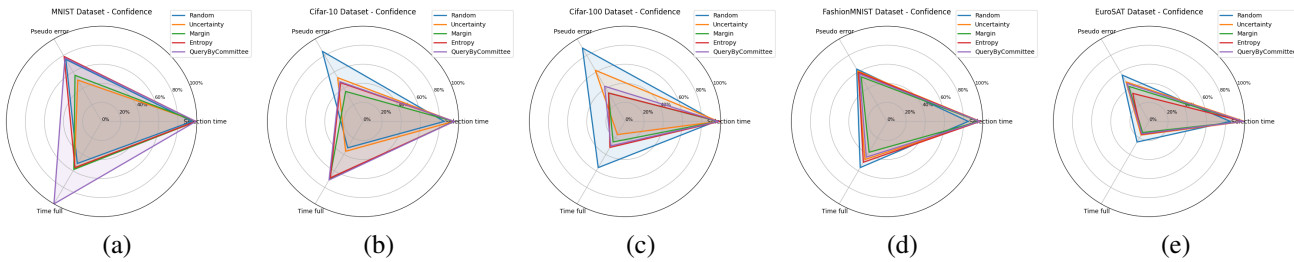


Figure 6. Confidence Pseudo - Strategy : Pseudo error, Selection Time and Time full Metrics for each dataset, presented in radar charts Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

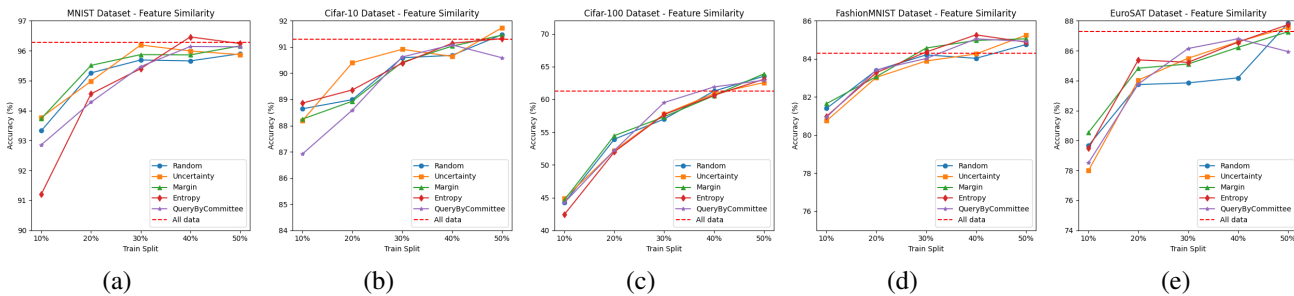


Figure 7. Feature Similarity Pseudo - Strategy : Accuracy per round comparison of Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

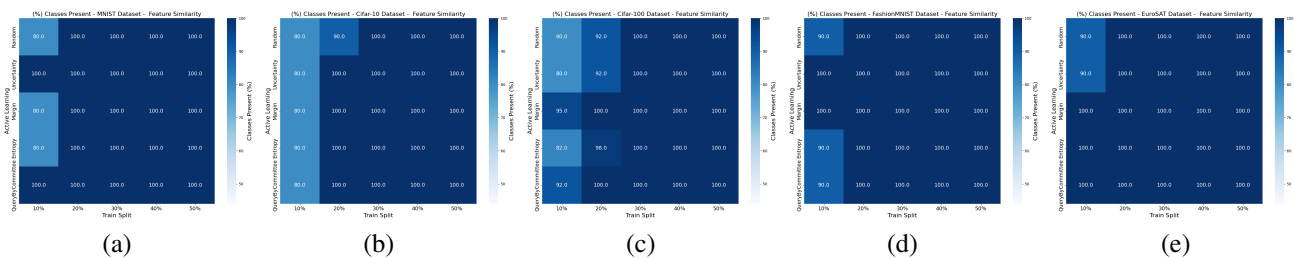


Figure 8. Feature Similarity Pseudo - Strategy : Percentage of classes present ($P_{classes}^i$) per iteration of Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

pseudo-labeled samples, allows for a significant maximization of the training process. Notably, our models trained under these conditions surpassed the performance in some scenarios using the entire dataset, including random sampling, highlighting the value of intelligence in data selection and the exploration of unlabeled data. We can conclude that the integration of active learning and pseudo-labeling represents a promising

path to enhance the development of classification models in limited-data domains, such as in various areas. The results presented not only validate the effectiveness of the approaches but also open new perspectives for future investigations. We hope that this work will inspire the development of new active learning methodologies, pseudo-labeling strategies, and the exploration of other applications of active learning and VLMs.

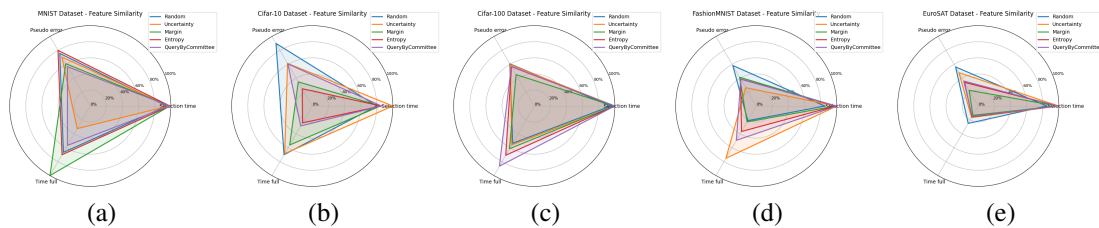


Figure 9. Feature Similarity Pseudo - Strategy : Pseudo error, Selection Time and Time full Metrics for each dataset, presented in radar charts Random, Uncertainty, Margin, Entropy and QueryByCommittee for each dataset: (a) Mnist, (b) Cifar-10, (c) Cifar-100, (d) FashionMNIST, and (e) EuroSAT.

Acknowledgements

This work was supported by the Eldorado Research Institute and the Federal University of Sao Carlos.

Author Contributions

Willian P. Amorim: Conceptualization, Methodology, Software, Writing – Original Draft. Priscila T. M. Saito: Supervision, Data Curation, Validation, Writing – Review & Editing. Cid Adinam Nogueira Santos: Project Administration, Funding Acquisition, Writing – Review & Editing.

References

- [1] RADFORD, A. et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, PMLR, p. 8748–8763, 2021.
- [2] ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, Nature Publishing Group, v. 542, n. 7639, p. 115–118, 2017.
- [3] WINDSOR, R. et al. Vision-language modelling for radiological imaging and reports in the low data regime. In: *Proceedings of Machine Learning Research*. [S.l.: s.n.], 2023. v. 227, p. 53–73.
- [4] NGUYEN, D. M. H. et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. [S.l.: s.n.], 2023.
- [5] ZHU, X. Semi-supervised learning literature survey. *Computer Sciences Technical Report 1530*, University of Wisconsin-Madison, 2005.
- [6] LEE, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *ICML Workshop on Challenges in Representation Learning*, 2013.
- [7] COHN, D. A.; GHAHRAMANI, Z.; JORDAN, M. I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, v. 4, p. 129–145, 1996.
- [8] ZHANG, J. et al. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 46, p. 5625–5644, 2023.
- [9] BANG, J.; AHN, S.; LEE, J.-G. Active prompt learning in vision language models. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2023.
- [10] LI, L. H. et al. VisualBert: A simple and performant baseline for vision and language. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 1492–1502.
- [11] LI, G. et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2020. v. 34, n. 07, p. 11316–11323.
- [12] CHEN, Y.-C. et al. Uniter: Universal image-text representation learning. In: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. Berlin, Heidelberg: Springer-Verlag, 2020. p. 104–120. ISBN 978-3-030-58576-1.
- [13] LU, J. et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2019. p. 13–23.
- [14] TAN, H. H.; BANSAL, M. Lxmert: Learning cross-modality encoder representations from transformers. In: *Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2019.
- [15] LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998.
- [16] KRIZHEVSKY, A.; HINTON, G. *Learning multiple layers of features from tiny images*. [S.l.], 2009.
- [17] XIAO, H.; RASUL, K.; VOLLGRAF, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In: *ICLR 2017 Workshop on Challenges in Representation Learning*. [S.l.: s.n.], 2017.
- [18] HELBER, P. et al. Eurosat: A novel dataset for deep learning in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, v. 57, n. 11, p. 9414–9426, 2019.