CLIP on low-resource vision

Trends Applications of Computer Vision
MSc in Artificial Intelligence Systems - University of Trento

Alessandro Lorenzi, Luca Cazzola, Omar Facchini

***Abstract* – The long-tailed data distribution pose a significant challenge in computer vision, a situation that is amplified in low-resource settings. In this paper we first categorize available literature which has shown appealing results facing such contexts. After that, some specific techniques being the bases for further studies are introduced in more detail. We conclude with an overview of the experiments we will conduct about emerging failures cases and class behaviours.**

I. INTRODUCTION

Deep learning has demonstrated significant success in numerous computer vision tasks; however, its dependency on large, balanced datasets poses a challenge for some real-world applications suffering from long-tailed data distributions phenomena. In such scenarios, a small subset of "head" classes is well represented, while the "tail" ones remain underrepresented. This imbalance leads to incomplete data representation, ambiguous decision boundaries, and poor generalization, particularly for tail classes.

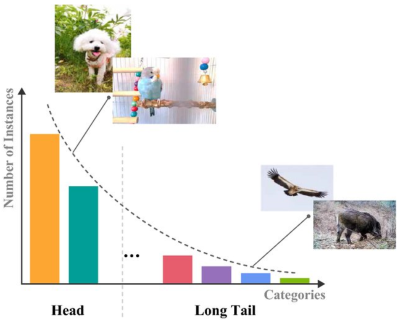

Fig. 1. The long-tailed data distribution

The ability to address this issue effectively is critical, especially in low-resource settings where acquiring additional data is infeasible or impractical.

II. BACKGROUND

In few-shot learning (FSL) the goal is to achieve performance using only a defined number of instances for each category. This situation is naturally intrinsic in low-resource contexts. Moreover, adopting FSL style of training encourages the model to pay more attention to tail classes, thanks to the fact that the distribution became balanced across all categories. This is still an effective approach even if the model has been pre-trained using a statistically different distribution of data, as in [1].

Going ahead with the various solutions proposed in literature to mitigate the cited problem [2], four main categories have been reported. Each class' methods are distinguished by their nature and strategy.

A. Data processing techniques

Working from the data perspective, the goal is to balance the dataset through distinct forms of samples processing. Over-sampling [3] duplicates tail class intances to increase their representation, while in under-sampling [4] the number of head examples is reduced. Both options have some drawbacks related to overfitting risk and to the potential loss of important information. Another processing technique is data augmentation [5] that, following several designs from deep learning generative oriented to more traditional ones, can be used to generate synthetic samples for tail classes. However, these artificially created samples may not always be realistic, lacking the fidelity needed to capture the true characteristics of the tail in real-world scenarios.

B. Cost-Sensitive Weighting

Cost-sensitive weighting methods assign different learning weights to classes or to individual samples to prioritize underrepresented categories. These approaches can be implemented at class level [6], in which the re-weighting strategy is typically based on inverse frequency of the classes, ensuring so that tail examples receive more focus during training. At instance level [7], weighting is instead applied on a per-sample basis, allowing for finer-grained control over the learning process. Even if these methods have shown some promising results in addressing the unbalancing problem, they are strongly related to both the dataset used and the considered task. This implies the need for careful tuning, generally leading to high resources cost and poor generalization.

C. Decoupling Methods

Decoupling methods [8] involve dividing the learning process into two distinct stages. In the first phase, a uniform sampling strategy is employed to learn general feature representations without taking into consideration the long-tailed problem. This helps avoiding model biases towards head classes early in the training process. In the second stage, class-balanced sampling is applied to ensure that the model handles data imbalancing more effectively. Despite their benefits and the increasing popularity in recent research, these techniques are not end-to-end solutions, as they disrupt the continuous flow of training in deep learning by introducing two separate steps. Additionally, the class-balanced sampling in the second stage presents similar challenges as the previous other methods, such as potential overfitting and inefficiencies in capturing the complexity of real-world data distributions.

D. Machine Learning-Based Methods

Machine learning-based methods often focus on leveraging specialized architectures, training paradigms and their combinations to better handle long-tailed data.

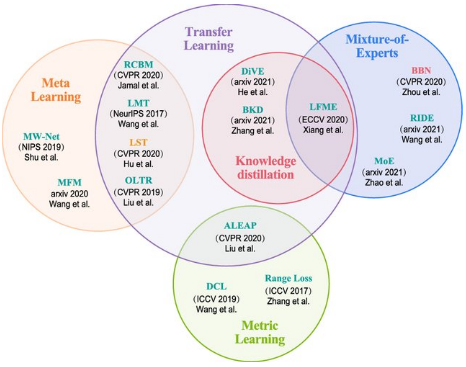

Fig. 2. Machine Learning-Based Methods

Metric learning [9] refines the feature space by mapping data points to an embedding space where distances between similar samples are minimized and those between dissimilar ones are maximized. This approach clarifies decision boundaries, particularly for tail classes, which typically suffer from ambiguous boundaries.

Transfer learning [10] can also be used to exploit all the knowledge acquired on head classes by a pre-trained model to guide classification for the tail instances. This is done through sharing model parameters across categories, resulting in the mitigation of the lack of information about the less populated classes.

Meta-learning [11] further extends this idea by training models in order to generalize across tasks and adapt quickly to new underrepresented labels using minimal data. This "learning to learn" concept allows an efficient training being able to transfer the classification capability for head classes to tail ones.

Mixture-of-experts [12] is a concept based on training multiple specialized networks, the "experts". Each one is responsible for a subset of the data, focused on a specific region, where it outperforms the others. This allow the model to better handle the imbalance among the classes. However, when building these architecture a critical part of the design is how to merge all experts answers or a subset of them.

Knowledge distillation [13], in turn, involves a teacher model with high performance guiding a simpler student model. The latter one learns to mirror the master, progressively improving its results. The goal is not just to have a smaller and better performing network with the same knowledge as the teacher, instead the outcome is supposed to be a more robust model, which is more sensitive to the tails.

Grouping techniques [14] attempt to segment the dataset into clusters and train the model separately on each group. While this may alleviate some class imbalance issues, the risk is to prevent the interaction of valuable knowledge across different groups, which could undermine the model's overall performance.

III. METHODS

Considering that the final purpose of this work is to test how different techniques perform on the long-tailed data distribution, a few methods have been selected, each one based on a different core idea and belonging to some of the defined categories. In particular, [15] [16] [17] are machine-learning based, while [18] falls into the data augmentations approaches.

A. Low-Rank Adaptation (LoRA)

In LoRA [15] the main intuition lies in learning low-rank matrix decomposition of parameter sets as efficient LoRA modules, while keeping original weights frozen.

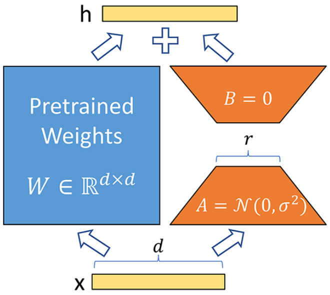
h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Fig. 3. LoRA module

When added to transformer layers [1], it is able to tune them efficiently, increasing performances on more specific tasks, while reducing the number of trainable parameters and preserving inference speed. This method is very suited for low-resource tasks as it allows for efficient adaptation without requiring extensive computational resources.

B. Bias-terms Fine-tuning (BitFit)

BitFit [16] is mainly about exposing knowledge already present in the model rather than learning new task-specific additional components. This is achieved by adjusting all the bias terms of the model, or a subset of them, while keeping other parameters frozen. This is another method solely relying on parameter tuning efficiency.

C. Meta-Adapter

Meta-Adapter [17] is designed to facilitate online adaptation with minimal examples. By taking a subset of support images, it learns a better alignment through cross-attention between the source and target feature representations.

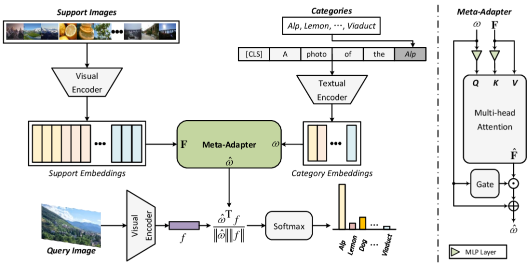

Fig. 4. Meta-Adapter module

This allows the extraction of valuable knowledge from few-shot samples, using a meta-learning style approach.

D. Label preserving breaking data augmentation

This method [18] augments datasets leveraging generative models capabilities, such as Stable Diffusion. Two types of images are produced from the original source. The generative process takes advantage of the diffusion model pipeline to perform less noise injection steps for the ”label-preserving” augmentations and more in the case of ”label-breaking” ones. Label-preserving images maintain the semantics of existing classes, while label-breaking ones introduce more diversity. The first are used in the supervised task loss and the latter ones are used to compute an additional term in an unsupervised way, using the contrastive paradigm.

$$L = L_{\text{task}} + \lambda L_{\text{label-breaking}}$$

$$L_{\text{label-breaking}} = -\frac{1}{N}\sum_{j}^{N} \log \frac{\exp\left(x_j'^{\mathrm{T}} x_j/\sigma\right)}{\sum_{i=1}^{N} \exp\left(x_j'^{\mathrm{T}} x_i/\sigma\right)}$$

A memory bank of maximum length N is stored, caching features of label breaking images. Original images and label preserving images come with label information and contribute to the L_{task}. Label breaking images are sampled 2 at a time and pushed close by the loss (x_j', x_j), while being put in contrast to the feature bank content.

IV. EXPERIMENTS

A. DATASETS

The tests are planned to be done on two different datasets: Eurosat [19] and Circuits [18], with the latter being an example of an extreme low-resource case.

Eurosat [19] is a dataset generally used as a benchmark for deep learning application in tasks of land use or land cover classification. It consists of satellite images divided into 10 classes of various environments varying between Vegetation places, Crop Lands, buildings and water-related locations. The size of this dataset is limited, but not restricted enough to be considered low-resources as it contains 27,000 labeled samples well balanced among the classes. This is still a challenging benchmark due to the nature of satellite images. Circuits is a newly presented dataset presented in [18]. The dataset consists of circuit diagram images belonging to 32 different classes such as “Audio amplifier”, “Relay” etc. The idea of it is to classify the schematics according to their function. Circuit-diagrams can be definitely considered a low-resource dataset, partially due to its nature and also given that sample are few ($\sim$ 1200). Another challenge present in this dataset lies in the nature of its content. Tiny differences in the electronic circuits representations could drastically change their function and therefore class. Having different layouts for the same operation might lead the model to overfit the specific design.

B. EVALUATION METRICS

The evaluation is planned to be done on many different metrics, in order to support out studies based mostly on failure cases analysis, as well as class behaviours.

- **Accuracy-related**: top-1, top-1 per class, confusion matrix.
- **Visualizations**: top-k correct predictions maximizing similarity, top-k incorrect predictions maximizing wrong-class similarity. Both settings are reproduced also per class and the relative attention maps is available for all visualizations.
- **Clustering-based**: 2D UMAP projection plots, Silhouette score, Adjusted Rand Index (ARI), V-measure. We consider as cluster assignments the model predictions and the actual target labels.
- **Improvements**: comparison of pair-wise different model settings, in which samples are wrongly classified by the first model and correctly labeled in the second one. Top-k couples are considered based on maximization of logits-entropy difference. The idea is to show examples which go through an important distribution shift with respect to the two settings.

C. EARLY RESULTS

Example of early results are shown in Fig. 5 and 6.

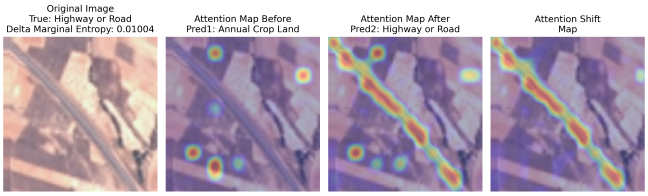

Fig. 5. Improvements example:
Zero-shot CLIP vs LoRA on Eurosat

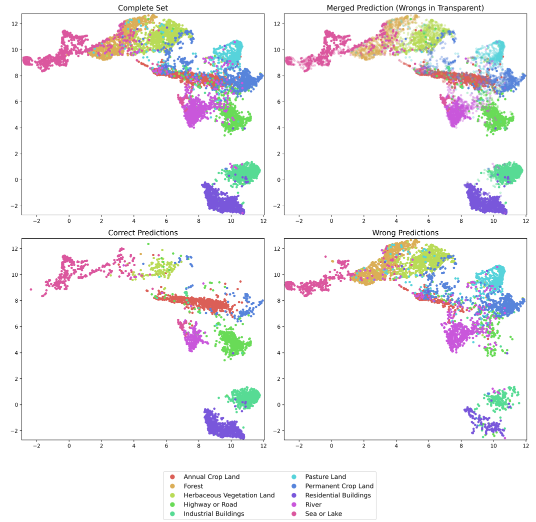

Fig. 6. Clustering-based example: BitFit on Eurosat

V. CONCLUSIONS

In this report we have first introduced the long-tailed data distribution problem, then we categorized and exposed some approaches attempting to mitigate those scenarios. To conclude, in the context of future work, we went more in depth in some architectures, particularly in the few-shot learning setting.

REFERENCES

[1] M. Zanella, I. B. Ayed, "Low-Rank Few-Shot Adaptation of Vision-Language Models", , 2024, URL: `https://arxiv.org/abs/2405.18541`, `2405.18541`.

[2] L. Yang, H. Jiang, Q. Song, J. Guo, "A Survey on Long-Tailed Visual Recognition", *International Journal of Computer Vision*, vol. 130, 07 2022, doi: 10.1007/s11263-022-01622-8.

[3] J. Kim, J. Jeong, J. Shin, "M2m: Imbalanced Classification via Major-to-Minor Translation", *in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13893–13902, 2020, doi:10.1109/CVPR42600.2020.01391.

[4] D. Devi, S. Biswas, B. Purkayastha, "Redundancy-driven modified Tomek-link based Undersampling: A Solution to Class Imbalance", *Pattern Recognition Letters*, 10 2016, doi:10.1016/j.patrec.2016.10.006.

[5] Y. Zang, C. Huang, C. C. Loy, "FASA: Feature Augmentation and Sampling Adaptation for Long-Tailed Instance Segmentation", , 02 2021, doi: 10.48550/arXiv.2102.12867.

[6] B. Li, Y. Yao, J. Tan, G. Zhang, F. Yu, J. Lu, Y. Luo, "Equalized Focal Loss for Dense Long-Tailed Object Detection", , 2022, URL: `https://arxiv.org/abs/2201.02593`, `2201.02593`.

[7] T.-I. Hsieh, E. Robb, H.-T. Chen, J.-B. Huang, "DropLoss for Long-Tail Instance Segmentation", vol. abs/2104.06402, 2021.

[8] C. Wang, S. Gao, C. Gao, P. Wang, W. Pei, L. Pan, Z. Xu, "Label-Aware Distribution Calibration for Long-tailed Classification", , 2021, URL: `https://arxiv.org/abs/2111.04901`, `2111.04901`.

[9] J. Cui, Z. Zhong, S. Liu, B. Yu, J. Jia, "Parametric Contrastive Learning", , 2021, URL: `https://arxiv.org/abs/2107.12028`, `2107.12028`.

[10] B. Liu, H. Li, H. Kang, G. Hua, N. Vasconcelos, "GistNet: a Geometric Structure Transfer Network for Long-Tailed Recognition", , 2021, URL: `https://arxiv.org/abs/2105.00131`, `2105.00131`.

[11] B. Li, Y. Liu, X. Wang, "Gradient Harmonized Single-Stage Detector", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8577–8584, Jul. 2019.

[12] X. Wang, L. Lian, Z. Miao, Z. Liu, S. X. Yu, "Long-tailed Recognition by Routing Diverse Distribution-Aware Experts", , 2022, URL: `https://arxiv.org/abs/2010.01809`, `2010.01809`.

[13] Y.-Y. He, J. Wu, X.-S. Wei, "Distilling Virtual Examples for Long-tailed Recognition", , 2021, URL: `https://arxiv.org/abs/2103.15042`, `2103.15042`.

[14] J. Wu, L. Song, T. Wang, Q. Zhang, J. Yuan, "Forest R-CNN: Large-Vocabulary Long-Tailed Object Detection and Instance Segmentation", , 2021, URL: `https://arxiv.org/abs/2008.05676`, `2008.05676`.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models", , 2021, URL: `https://arxiv.org/abs/2106.09685`, `2106.09685`.

[16] E. B. Zaken, S. Ravfogel, Y. Goldberg, "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models", , 2022, URL: `https://arxiv.org/abs/2106.10199`, `2106.10199`.

[17] C. Cheng, L. Song, R. Xue, H. Wang, H. Sun, Y. Ge, Y. Shan, "Meta-Adapter: An Online Few-shot Learner for Vision-Language Model", , 2024, URL: `https://arxiv.org/abs/2311.03774`, `2311.03774`.

[18] Y. Zhang, H. Doughty, C. G. M. Snoek, "Low-Resource Vision Challenges for Foundation Models", , 2024, URL: `https://arxiv.org/abs/2401.04716`, `2401.04716`.

[19] P. Helber, B. Bischke, A. Dengel, D. Borth, "Introducing EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification", *in IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 204–207, IEEE, 2018.