Collaborative Classification for Object Labeling on Expansible Datasets

Bruno Padilha[1], **João E. Ferreira**[1]

[1]Institute of Mathematics and Statistics - University of São Paulo (IME-USP)
São Paulo – SP – Brazil

brunopadilha@usp.br, jef@ime.usp.br

Abstract. *Streaming applications in video monitoring networks generate datasets that are continuously expanding in terms of data amount and sources. Thus, given the sheer amount of data in these scenarios, one big and fundamental challenge is how to reliably automate data annotation. In this work, we propose a novel active learning strategy based on multi-model collaboration able to self-annotate training data providing only a small initial subset of human verified labels, towards incremental model improvement and distribution shifts adaptation. To validate our approach, we collected approximately 50,000 hours of video data sourced from 193 security cameras from University of São Paulo Monitoring System (USP-EMS) during the years 2021-2023, totaling 7.3TB of raw data. For experimental purposes, this work is focused on identification of pedestrians, cyclists and motorcyclists resulting in 3.5M unique objects labeled with accuracy between 92% to 96% for all evaluated cameras.*

1. Introduction

State-of-the-art deep learning models for image classification rely on large volumes of annotated data (e.g. [Kirillov et al. 2023]). Once upon a time, obtaining data for machine learning was costly and difficult to come by due to technology restrictions in availability of sensors (i.e. cameras, social networks, signal detectors, etc...), data storage and processing power. Nowadays, data is cheaper, easier to come by and being produced at an accelerating pace. On the other hand, annotating data for supervised learning remain expensive once human generated labels are still pervasive in many successful training strategies. In spite of recent advancements in object tracking algorithms, pre-trained models and other tools that can assist humans to speed up data annotation, the cost is still high and can increase faster than linear with the dataset size.

One low-cost way to leverage pre-trained models is known as Transfer Learning (TL) and consists in fine-tuning a pre-trained model with a much smaller dataset of a target domain. It is feasible providing the two domains (original and targeted) share good amounts of general object attributes and data distributions are not too far apart. However, it is harder to find such large datasets for niche domains that would allow us to apply TL in more domain specific downstream tasks (e.g. medical images, manufacturing quality control, agricultural applications). Alternatively, another approach to alleviate the burden of annotating a new dataset is known as Active Learning (AL) [Ren et al. 2021b]. In AL, the main concern is how to attain the best possible performance from a model with a minimal amount of annotated data. In other words, data can labeled in small amounts to

train a model in an iterative manner and the previous acquired knowledge is leveraged to devise a query strategy to label new samples.

In this work, we present a novel method based on the combination Transfer Learning with Active Learning to reliably annotate large amount of objects in video data in a semi-automated manner. It is composed by teams of binary classifiers whose decisions are made **collaboratively** by voting and consensus policies. Initially, a small subset (e.g 1000 samples per class) of a raw dataset is randomly sampled for human verification to train weak classifiers. Data is partitioned in a one vs. rest (OvR) fashion to minimize the odds of teams yielding arbitrary high confidence outputs for far out-of-distribution samples [Hein et al. 2019]. This strategy allows these classifiers to self-annotate data as a team relying on partial acquired knowledge in order to incrementally expand the training set. Our solution should not be confused with ensemble learning [Zhang and Ma 2012]. Contrary to the later, member of our classification teams evolve from weak classifiers at training phase to full fledge independent models.

We demonstrate the effectiveness of our method with classification experiments on real data for the classes *cyclist*, *motorcyclist* and *pedestrian*. The classification data contains $\sim$3.5M unique object samples extracted from 50.000h of raw video footage from 193 cameras from USP-EMS (Electronic Monitoring System at University of São Paulo) [Ferreira et al. 2018] collected between the years of 2021-2023. In all experiments, the teams of experts were able to consistently learn the cameras distributions, reaching up to 96% in accuracy, while using only a few thousand of automatically labeled samples per class. Intending to better understand the challenges of learning with AL and the teams of experts, we opted to approach it as a classification problem leaving added complexities of detection and segmentation problems for future works.

2. Related Work

2.1. Expansible Dataset

The new dataset we are going to build is continuously sourced by video footage from security cameras monitoring open public areas in the dependencies of the University of São Paulo (USP). Nowadays, it is common to find application domains generating data streams, subjecting models trained on static datasets to knowledge obsolescence [Zhang et al. 2022]. We named *Expansible Datasets* this emerging category of datasets that are continuously expanding with novel data.

2.2. Active Learning

The main concept behind Active Learning (AL) is to start training a model with only a small fraction of reliable (i.e. human verified) labeled data. Then, this partial learned knowledge can be employed to discover which unlabeled samples will contribute the most to improve the model this time. The process repeats until the model meets some performance criteria. Another key concept is how to apply a model to query the unlabeled set. According to [Ren et al. 2021b], AL strategies can be categorized into membership query synthesis, stream sampling and pool sampling. In the context of deep learning, the first one is usually related to sample generation, for example with GANs or VAEs models, and request it to be human labeled. The second one is suitable for storage and computing limited devices in which is there is no access to at least a sizable portion of the unlabeled

set. In the last one, pool sampling, model knowledge is used to rank unlabeled data based either on sample diversity [Hwang et al. 2024].

2.3. Out-of-distribution detection

Out of distribution (OOD) detection is the ability of a model to recognize data samples that deviate from data distribution of learned representations. According to [Winkens et al. 2020], OOD detection problems are more challenging when OOD samples are near the in-distribution ones (near-OOD) than when they appear farther away (far-OOD). Expansible datasets contemplates both scenarios once new *unlabeled samples* can contain novel knowledge (near-OOD), valuable to improve classification, or simply detrimental noise (far-OOD). One simple approach the far-OOD case is a method know as the Mahalanobis Distance (MD), a function to compute the distance of a point to a known distribution. Several authors ([Ren et al. 2021a], [Fort et al. 2021]) have been trying to improve the MD method in the near-OOD case. However, these proposals rely on large and consolidate datasets with reliable annotations (e.g. CIFAR-10, CIFAR-100, ImageNet-21k), which initially is not available when building a new dataset.

2.4. Overconfidence and Uncertainty Estimation

Most modern deep learning models are based on the softmax to compute probabilities and the cross-entropy as the loss function. However, these probabilities can be overestimated and do not represent true likelihood [Guo et al. 2017]. Moreover, training data with one-hot labels may lead to cross-entropy overfitting to labels before it actually overfit to data. Both issue lead to overconfident models and hinders uncertainty estimation. As mitigating measures, [Kristiadi et al. 2020] proposes an adversarial training technique to enforce low confidence for far out-of-distribution data. Label smoothing, which we have employed in our method, has been proven [Zhang et al. 2021] to be a effective regularization technique to soften one-hot labels and mitigate overconfidence.

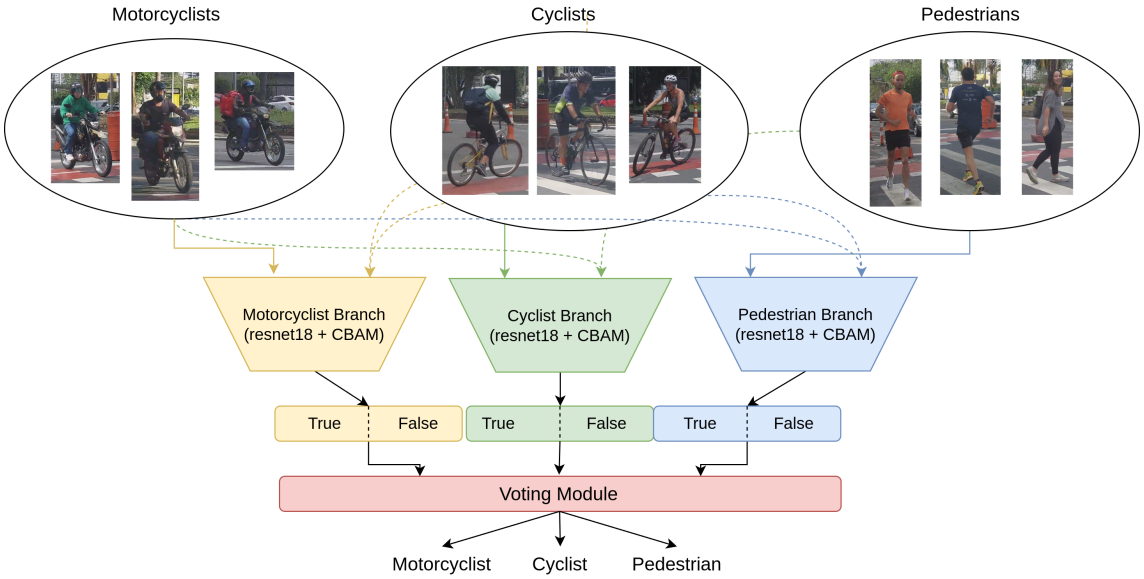
3. Proposed Method

Small datasets sampled from real world sources, lets say about 1000 samples per class, even in the absence of label or sample noise are usually insufficient to properly learn a domain distribution. However, this limited dataset do contain some knowledge to train weak classifiers. Furthermore, in case we can segment this dataset per data source (i.e per camera), we can train a team of n independent binary weak classifiers (one for each class) that can collaboratively reach a voting-based consensus to identify near-OOD samples, which we empirically show that are the ones containing novel information regarding the distribution of this dataset segment. A batch of these near-OOD samples, automatically labeled by the team, can be selected based on the team joint confidence to increment the training set and expand the model knowledge. Because we start with weak classifiers, we increase the training set in small increments to avoid absorbing too much noise. After only a few rounds of increment-and-train, we have a *team of experts* for that distribution.

3.1. Model architecture

Figure 1 depicts our proposed architecture for a team of classifiers. All members of this team, also called *branches*, are ResNet-18 paired with CBAM attention layers. Branches are binary classifiers specialized in a single class, that is, they individually decide if a

Figure 1. Classification team architecture



given sample belongs (true) or do not belong (false) to that class. Input data for training is split in a One-vs-Rest (OvR) fashion where, for each branch, the true class contain only samples of a specific class (e.g. motorcyclist) and the false class is composed of a combination of samples from all other classes (e.g cyclist + pedestrians). In our experiments, this training data arrangement has been demonstrated to be a reliable way to approximate the pseudo-class *I don't know* when decisions are made collaboratively by the members of a classification team. Moreover, in order to mitigate the overconfidence problem that may occur when training ResNets with piece-wise linear activation functions (e.g. Relu and variants), models are trained with label smoothing. In summary, we have the following hyper-parameters:

- Max epochs: 30
- Batch sizes: Train = 64, Test = 16
- 5-fold validation
- Kaiming weights initialization
- Weighted Cross-Entropy with label smoothing as the loss function
- Gradients calculated with SGD (learning rate = 0.1, momentum = 0.9, weight_decay = $5 * 10^{-4}$)
- Cosine Annealing learning rate scheduler

3.2. Team consensus

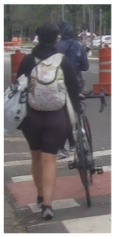
When in evaluation mode, individual decisions are combined in the voting module. One simple yet effective voting strategy is the *consensus*, meaning all branches must agree on one class. For example, Figure 2 illustrates the results of the evaluation of a picture containing a motorcyclist for which the "motorcyclist" branch voted true while the other two branches, "cyclist" and "pedestrian", voted false, thus reaching an agreement for classifying this image as motorcyclist. On the other hand, Figure 3 illustrates a case of no consensus for which both the "cyclist" and the "pedestrian" branches voted true. In

this case, the team as a whole could not decide and the verdict is ”I don’t know”. The confidences of each branch will be used as thresholds to decide what images should be considered for expanding the training set for the next iteration.

	Branch		
	Motorcyclist	Cyclist	Pedestrian
Vote	True	False	False
Confidence	96.7%	93.1%	95.1%

Prediction: Motorcyclist

Figure 2. Evaluation of a picture containing a motorcyclist for which the ”motorcyclist” branch voted true while the other two branches, ”cyclist” and ”pedestrian”, voted false, thus reaching an agreement



	Branch		
	Motorcyclist	Cyclist	Pedestrian
Vote	False	True	True
Confidence	92.3%	73.8%	85.5%

Prediction: *"I don't know"*

Figure 3. Team was unable to reach a consensus and the final verdict is ”I don’t know”

4. Experiments

The main source of data for the following experiments is USP-EMS (Electronic Monitoring System) [Ferreira et al. 2018]. It contains hundreds of security cameras to monitor USP dependencies in eight campuses in the state of São Paulo, Brazil. The footage used in this work was collected during the years of 2021 and 2022, spanning varying seasons, weather and times of the day. Each source video is one hour long. These cameras were hand picked in close collaboration with campus security department to reflect regions and times of biggest traffic movement, cyclist concentration, street crossing and some intersections prone to intercurrences. In order to avoid manual annotating data from scratch, we leverage transfer learning by feeding raw video data to a pre-trained object detector and tracker for objects cropping. For this purpose, the most suitable tool we found was a combination of YOLOv8 pre-trained on COCO dataset, the strongSort object tracking algorithm [Du et al. 2023] and our custom cyclist and motorcyclist detection algorithm [Nardi et al. 2022].

4.1. Incremental Learning

Once the initial dataset is human validated, we proceed to the first round of incremental learning. Table 1 presents the evolution of branches training after six rounds of increment-and-train. The dataset is split into 80% for training and 20% for validation. Evaluation metric is the recall for the individual *voting classes* True and False. The numbers i1 ... i5 represent the new set of thousand images per class as selected by the team through the

consensus mechanism. For the selection strategy, we evaluate on the remaining samples in the raw dataset for that specific camera and sort the output of the consensus by confidence, considering only the top 25% values (the upper quartile) as candidates. Among the candidates, selection of the next thousand samples, is random once it demonstrates to be beneficial to mitigate inductive biases introduced by human selecting samples in step i0, which if is not addressed, may propagate to subsequent iterations and potentially degenerate some models.

Table 1. Recall results of Incremental learning for six rounds on camera S10-08 (Praça R) with a three members team (motorcyclist, cyclist and pedestrian).

Camera S10-08 (Praça R): Branches Recall

	Motorcyclist		Cyclist		Pedestrian	
	true	false	true	false	true	false
i0 (human)	91.5%	92.2%	88%	92%	88%	94.2%
i0+i1	93.5%	94%	93.7%	95.1%	92.2%	94.3%
i0+i1+i2	93.6%	90.2%	94.8%	94.9%	93.1%	92%
i0+i1+i2+i3	97.6%	93.4%	96.6%	95%	91.1%	94.1%
i0+i1+i2+i3+i4	95.6%	94%	96%	96.7%	93%	96.9%
i0+i1+i2+i3+i4+i5	96.3%	95.1%	95.8%	96.9%	94.6%	96.4%

4.2. Conclusions, Limitations and Assumptions

This work was initially developed to address a real-world demand to produce reliable annotations for non-stop growing datasets of objects extracted from security cameras in USP-EMS. In this scenario, we are able to define individual cameras as local domains, thus producing local datasets. The concept of near/far OOD in this limited view of the world, although based on real-word data, is more self-behaved than applying the same concept without imposing any restriction on global data (from all cameras sources). In order to overcome this limitation, we are currently working on an improved version of our approach based on concepts and techniques proposed in some of the works presented in Section 2.

Acknowledgment

This work was supported by The São Paulo Research Foundation, FAPESP (grant number. 2020/06950-4) and The National Council for Scientific and Technological Development (CNPq) - CNPq Research Productivity Scholarship Program.

References

Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., and Meng, H. (2023). Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*.

Ferreira, J. E., Antônio Visintin, J., Okamoto, J., Cesar Bernardes, M., Paterlini, A., Roque, A. C., and Ramalho Miguel, M. (2018). Integrating the university of são paulo

security mobile app to the electronic monitoring system. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1377–1386. IEEE.

Fort, S., Ren, J., and Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hein, M., Andriushchenko, M., and Bitterwolf, J. (2019). Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50.

Hwang, Y., Jo, W., Hong, J., and Choi, Y. (2024). Overcoming overconfidence for active learning. *IEEE Access*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Kristiadi, A., Hein, M., and Hennig, P. (2020). Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pages 5436–5446. PMLR.

Nardi, E., Padilha, B., Kamaura, L., and Ferreira, J. (2022). Openimages cyclists: Expandindo a generalização na detecção de ciclistas em câmeras de segurança. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 229–240, Porto Alegre, RS, Brasil. SBC.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. (2021a). A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021b). A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al. (2020). Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*.

Zhang, C. and Ma, Y. (2012). *Ensemble machine learning*, volume 144. Springer.

Zhang, C.-B., Jiang, P.-T., Hou, Q., Wei, Y., Han, Q., Li, Z., and Cheng, M.-M. (2021). Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30:5984–5996.

Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., and Liu, H. (2022). Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4910–4920.