HUMANS ARE POOR FEW-SHOT CLASSIFIERS FOR SENTINEL-2 LAND COVER

Marc Rußwurm[1], *Sherrie Wang*[2], *Devis Tuia*[1]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] University of California, Berkeley, USA

ABSTRACT

Learning to predict accurately from a few data samples is a central challenge in modern data-hungry machine learning. On natural images, human vision typically outperforms deep learning approaches on few-shot learning. However, we hypothesize that aerial and satellite images are more challenging to the human eye. This applies particularly when the image resolution is comparatively low, as with the 10m ground sampling distance of Sentinel-2. In this study, we benchmark model-agnostic meta-learning (MAML) algorithms against human participants on few-shot land cover classification with Sentinel-2 imagery on the Sen12MS dataset. We find that categorization of land cover from globally distributed regions is a difficult task for the participants, who classified the given images less accurately than the MAML-trained model and with a highly variable success rate. This suggest that hand-labeling land cover directly on Sentinel-2 imagery is not optimal when tackling a new land cover classification problem. Labeling only a few images and employing a trained meta-learning model to this task may lead to more accurate and consistent solutions compared to hand labeling by multiple individuals.

Index Terms— Few-shot Land Cover Classification, Model-Agnostic Meta Learning; Sentinel-2; Participant Survey; Visual Photointerpretation; Human Perception.

1. INTRODUCTION

Human learning typically requires few experiences to adapt to new unseen situations. We do this by leveraging prior knowledge learned throughout our lifetime for a new task. By contrast, data-driven machine learning models are typically trained from scratch and require large annotated datasets to learn suitable parameters for feature extraction and classification. The question of *how to transfer* any existing prior knowledge to new unseen tasks in machine learning remains an active field of research [1] and a highly relevant question in Earth observation [2, 3]. One effective strategy for transfer learning is meta-learning [4, 5], where an algorithm learns from a collection of source ("meta-train") tasks to adapt a model to new unseen ("meta-test") tasks. Data of each task is randomly split into a "support" partition to adapt the model and a "query" partition to measure the performance on the task. In few-shot learning terminology, a method that makes predictions after seeing k examples of n classes is said to perform n-way, k-shot classification. Hence, a trained few-shot model can learn to classify a new set of classes from an unseen region with only k examples per class. This makes few-shot learning models particularly useful in regions where labeled data is scarce and only a few hand-collected samples are available. As an example, consider the task of classifying images into two classes, "Grassland" and "Cropland" in Kenya without any existing ground truth on-site. Practitioners typically hand-label comparatively large datasets and potentially split the effort between multiple labeling individuals to generate sufficiently large datasets to train a deep learning model from scratch. Employing a few-shot meta-learning model that was trained on land cover classification tasks, for instance, from Germany with other classes, can transfer the knowledge from these related tasks and reduce the number of required labeled images. Meta-learning approaches have proven accurate in Earth observation [6], but the question of whether these approaches perform the same (or better) than humans visually classifying land cover in unseen regions remains unanswered. Nonetheless, it is a very relevant question, since meta-learning for few-shot learning only requires a few

Fig. 1: The web interface that the participants used to assign a single query image (top) to one of four categories (bottom).

labeled examples per task, and could therefore achieve a significant speedup with respect to human visual classification. Human classification is often considered more accurate, but it is also prone to errors and uncertainties [7], which become more prominent when the resolution decreases and creates the necessity to work on ensembles of human users confirming each others' decisions [8]. Therefore, in this study, we run an experiment with volunteer photo-interpreters and compare the accuracy of their labeling with the outcome of several meta-learning algorithms.

2. FEW-SHOT META LEARNING

Few-shot meta-learning methods can be grouped into two categories. *Gradient-based methods*, such as model-agnostic meta-learning (MAML) [9] and MAML++ [10], strive to find a model initialization that adapts quickly to new tasks. Training a model progresses in two nested loops. The inner loop fine-tunes a differentiable model on each task with few support samples, while the outer loop updates the initialization parameters based on the query samples. These outer parameters then capture the prior knowledge from all meta-training tasks and help the model to adapt to unseen meta-test tasks. *Metric-based methods*, such as prototypical networks [11], learn a model that embeds the support and query samples into a common feature space. The feature representations of the support samples of each class are averaged to a class prototype. The query samples are then assigned to the closest class prototype.

3. EXPERIMENTAL SETUP

3.1. Data and Few-Shot Task Definition

We define tasks from land cover classification problems with the Sen12MS [12] dataset. Sen12MS provides Sentinel-2 and Sentinel-1 images alongside a simplified International Geosphere-Biosphere Programme (IGBP) land cover label scheme of 10 classes (Forest, Shrubland, Savanna, Grassland, Wetlands, Croplands, Urban/Builtup, Snow/Ice, Barren, Water) across 125 globally distributed regions. A subset of 75 regions is used for meta-train tasks while 25 regions are used for meta-validation and meta-test datasets each. Each region is split into overlapping 256×256 pixels tiles which we split into non-overlapping patches and assigned the label of the most common land cover to this patch [6]. After this split, a dataset sample becomes a 15-dimensional, 128×128 pixels patch (composed of a stack of the 13 Sentinel-2-bands and the two Sentinel-1 polarizations) and one associated label. In the case of mixed land cover within one tile, the class of the most frequent land cover is assigned to this patch. A $k = 2$-shot $n = 4$-way task is sampled by first choosing four (out of ten) random classes of each region and then four ($2k$) examples per class. This dataset is then divided into support and query partitions with eight images (nk) each.

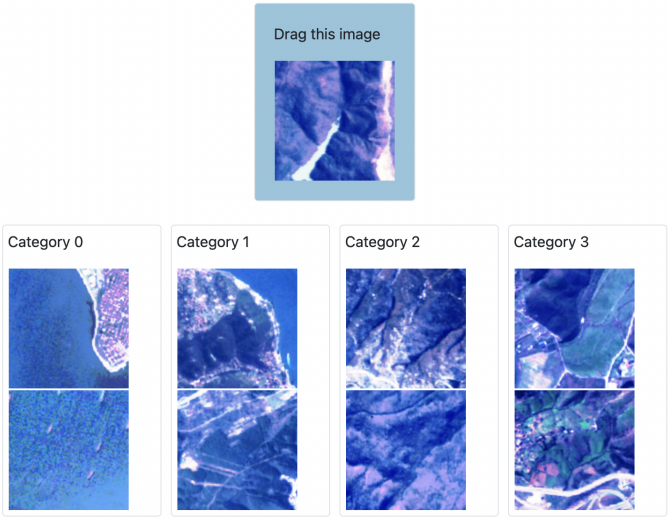

Fig. 2: One task posed to the participants. The query image top needs to be drawn to the corresponding category. In this task, the category 2 is correct which is not a trivial task for visual interpretation.

3.2. Model Design and Training

We compare several few shot meta-learning models to the human scores: MAML [9], MAML++ [10] and ProtoNet [11]. All few-shot methods used a convolutional classification model with 6 layers, where each layer consists of a 3×3 convolution with stride 1 followed by 2D Batch Normalization, a ReLU activation function, and a 2D pooling layer of stride 2. The first layer projects the 15 input bands to 64 hidden dimensions, while a final linear decision layer projects the features to 10 classes.

3.3. Participant Study

Selection. Twenty-one individuals participated in the study. They were all students in computer science and remote sensing, at the MSc and Ph.D. levels. The participants were familiar with computational Earth observation, but not with visual photo interpretation. Also, none of the participants was familiar with the Sen12MS dataset [12].

Interface. The participants used a web application (shown in Fig. 1) and were asked to classify 50 images. Each image was picked from a different 4-way 2-shot task. Each task consisted of four random land cover classes from one Sen12MS test region. Each class was represented by two examples that correspond to the support set of a 4-way 2-shot task to familiarize the participants with the representation of each class. In Fig. 2, we show one task to illustrate the problem. A single

	training algorithm	accuracy	κ
Meta-Learning	MAML [9]	0.81	0.74
	MAML++ [10]	0.81	0.75
	Fo-MAML [9]	0.81	0.75
	ProtoNet [11]	0.75	0.66
Participants	best individual	0.77	0.69
	best 25%	0.74	0.65
	median	0.60	0.47
	worst 25%	0.48	0.30
comparison	random guess	0.25	0

Table 1: Model comparison on few-shot image classification on the Sen12MS dataset

image from the query set was shown at the top, and the participant was asked to drag this image to the category that they deemed most similar. After each assignment, visual feedback of whether the classification was correct was given to the participants.

Survey Post-processing. Since participants were unfamiliar with the photo interpretation task and with the web interface itself, we expected a certain number of accidental or involuntary inputs. To address these, we discarded the first five classifications of each participant to ensure that the participants were familiar with the interface. Additionally, we removed assignments that were faster than 1.5 seconds or took longer than 20 seconds. These parameters were determined before inviting the participants. All entries remaining after these post-processing steps were used in the results presented in the next section.

4. RESULTS

In this section, we first compare the accuracy of machine learning models to that of the human participants. We then study the variance inaccuracy within the cohort of participants and correlate classification performance to decision duration.

4.1. Accuracy Comparison

Numerical results are reported in Table 1: gradient-based meta-learning methods MAML and MAML++ led to the highest accuracy of 81% compared to the one obtained by the five best participants (top quarter) of 74%. ProtoNet was slightly less accurate than MAML-based methods and roughly on par with the best human performer (but better than the human cohort as a whole).

4.2. Participant Variance

In Table 1, we observed a strong variance in the classification accuracy between participants, where the median accuracy was only 60% compared to the best participant, who classified the tasks at 77%. This relationship is shown on a per-participant basis in Fig. 3, where we can observe that the three best individuals predicted the samples at comparatively similar accuracy while the accuracy of other participants varied drastically between 42% and 69%.

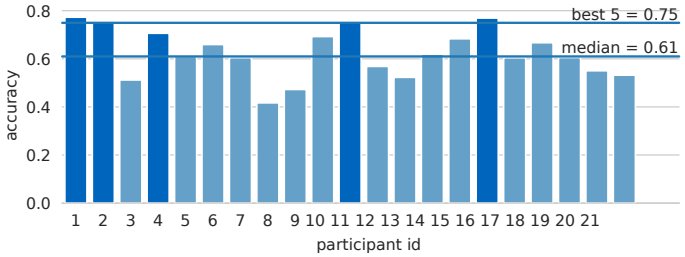

Fig. 3: Variance in accuracy for participants

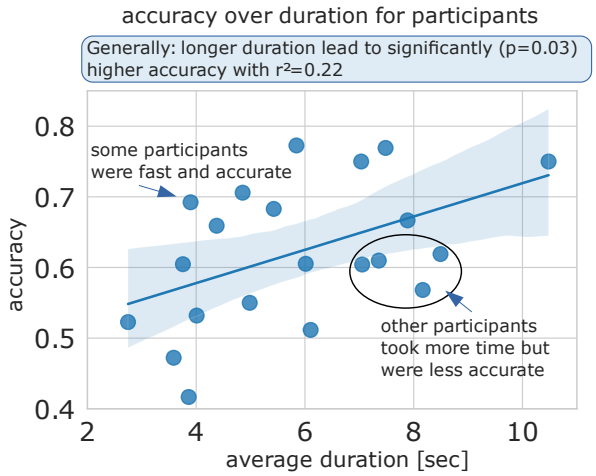

Fig. 4: average duration and accuracy per participant

4.3. Classification Duration

In Fig. 4, we study the relationship between assignment duration and accuracy for the participants. We find a significant ($p = 0.03$) but weak correlation ($r^2 = 0.22$) between average duration and accuracy. As one would expect, a more careful and thus longer class assignment duration leads to a more accurate prediction. Still, we note that the worst participant with 42% accuracy and an average assignment duration of 3.86s took one second longer than the fastest participant with $2.75s$ and 52% accuracy. The most accurate three participants took between 5.84 and 7.48 seconds per assignment, while a group of other participants took more time between $8.49s$ and $7.05s$ but assigned the classes at a lower accuracy between 57% and 62%. One participant achieved a remarkable accuracy of 69% with a decision duration of only 3.90 seconds while three participants required 8 seconds and were assigned the tasks only at a 55% to 65% accuracy range

5. DISCUSSION

In this study, we compared the accuracy of few-shot meta-learning methods to human participants for Sentinel-2 land cover classification. While we found that gradient-based meta-learning methods on multi-spectral images outperformed the best human participant in this study, we need to consider some caveats in the direct comparison of the machine and human performance.

Humans and models have access to different data (and experiences). First, all machine learning models had access to thirteen multi-spectral Sentinel-2 bands and two Sentinel-1 polarizations. In comparison, the participants could only see a histogram-normalized RGB representation. On the other side, the participants can draw from a potentially large knowledge-base of photo-interpretation experience while the meta-learning models were trained from scratch on a Sen12MS meta-train set. This could, up to some extent, also explain the large variability in accuracy and speed between subjects.

Models predicted IGBP categories. Second, we kept the label spaces fixed to ten classes for the MAML models, as the overall ten IGBP classes did not change between the tasks. Hence, the MAML model could predict any of ten land cover classes.In other words, the meta-learning models had the advantage of consistency of the definition of the label across tasks. This was not possible for human participants who had to learn the representation of each category solely from the provided support images of four randomly chosen classes.

Single query image for participants. The few-shot models predicted the entire query partition of each task. In our 2-shot 4-way setup, this means 8 images of the same region. On the contrary, human participants were asked to assign only one query image to one of the four randomly selected classes. We decided to proceed like this to maximize the number of regions seen by each participant, but this might have made the task harder for the participants but not for the model that treats all images independently.

6. CONCLUSION

The good accuracies of the few-shot machine learning models on this dataset showed that existing few-shot meta-learning algorithms outperform human vision in this area. The models could exploit all provided spectral bands and perform inference in fractions of a second while, generally, a longer assignment duration led to more accurate classifications by the human participants. If we consider land use/land cover classification in new areas as the final goal, we observe that labeling directly at Sentinel resolution is a challenging task for human photo-interpreters. Easing the task by considering very-high-resolution imagery is a suitable alternative, but requires purchasing potentially expensive satellite data. Overall, the results of this study suggest that employing a few-shot meta-learning model that is trained on land cover classification problems from other regions is a promising idea, which after fine-tuning on a few samples led to more consistent and accurate results than those obtained by the cohort of human interpreters.

7. REFERENCES

[1] Qiang Yang, Yu Zhang, Wenyuan Dai, and Sinno Jialin Pan, *Transfer Learning*, Cambridge University Press, 2020.

[2] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone, "Recent advances in domain adaptation for the classification of remote sensing data," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[3] Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein (Editors), *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*, Wiley & Sons, 2021.

[4] Jürgen Schmidhuber, *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*, Ph.D. thesis, Technische Universität München, 1987.

[5] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier, *Learning a synaptic learning rule*, Citeseer, 1990.

[6] Marc Rußwurm, Sherrie Wang, Marco Körner, and David Lobell, "Meta-learning for few-shot land cover classification," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, 2020, pp. 200–201.

[7] Devis Tuia and Jordi Muñoz-Marí, "Learning user's confidence for active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 872–880, 2013.

[8] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4205–4230, 2021.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[10] Antreas Antoniou, Harrison Edwards, and Amos J. Storkey, "How to train your MAML," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2019, OpenReview.net.

[11] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[12] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu, "Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019, vol. IV-2/W7, pp. 153–160.