Wildfire Prediction Challenge

Petro Shulzhenko

Télécom Paris, X2021

I. Introduction

This work is based on a more complicated version of Wildfire Prediction Dataset (Satellite Images) Challenge https://www.kaggle.com/datasets/abdelghaniaaba/wildfire-prediction-dataset/data, focusing on the binary classification of satellite images into wildfire and no-wildfire categories. The aforementioned complication consists in prohibiting the use of annotations of the original training dataset, which opens up interesting possibilities to research transfer, active, semi-supervised and unsupervised learning methods. Another goal of the work was to strive for independent implementation of the approaches and instruments used. Thus, all methods, pipelines and utilities, with the exception of only one of the used pre-trained models, were written in PyTorch, without the use of third-party projects. The project code can be found here https://github.com/ShulzhenkoPetr/SemiSupervisedLearning_Wildfire_Challenge.git.

II. Dataset

The original dataset is split into three parts: train - 70%, validation - 15%, test - 15% of the total number of images. By removing annotations for the training dataset we are left with: 6300 annotated images of the original validation dataset, which we split into training and validation in the proportion 85-15; 30250 unannotated images of the original train set and 6300 images of the test set. Thus, the number of annotated images for the train is 5355 and for the validation 945. The following features of the dataset are of interest:

- In both the original and used dataset, there is a slight class imbalance towards images containing wildfires, in a proportion of about 53 to 47.
- The dataset contains both forest images and images of, for example, cities. Which to some extent may simplify the task for the DL-based model and raise the question of what exactly the model learns and uses for classification.
- Satellite images are quite different from classical datasets such as ImageNet [2], CIFAR etc., which raises the question of the relevance of features learned on those datasets in application to this problem.
- The image statistics in the dataset are also different from ImageNet. For the dataset in consideration (based on 30 thousand images): Mean: [0.297 0.348 0.253] Std: [0.195 0.163 0.168]. In turn, it is known that the optimal constants for ImageNet are: Mean=[0.485, 0.456, 0.406], Std=[0.229, 0.224, 0.225]
- Images are provided in RGB format with dimensions 350x350. As part of the project all pipelines resize images to size 224x224 (without Crop because many images contain fire traces exactly on the edges) to use models pre-trained on ImageNet, and also because for the used models this size is optimal for their receptive fields, according to recommendations from the paper.

III. Supervised approach

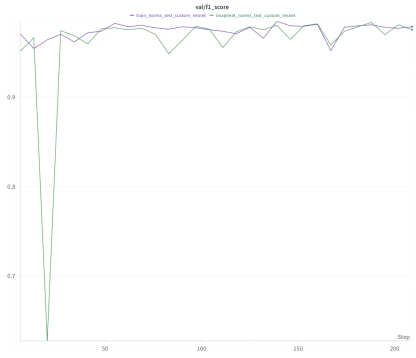
First, supervised methods were considered to obtain baselines. VGG-like [14] (the name is used for convenience to describe a model with standard blocks without skip-connections, inception blocks etc.) model and the custom ResNet [7] model (3M params) were implemented and trained from scratch. Additionally, pre-trained on ImageNet ResNet-18 model was fine-tuned with different numbers of frozen layers. Due to the small size of the train dataset, different sets of standard augmentations were used: Flips, Rotations, ColorJitter. Various learning techniques have also been used, including WarmUp scheduler for learning rate, which may have an important role in combination with the Adam optimizer that was used.

Experiments with different image normalization constants did not show their significant influence on the training results, due to the sufficiently large size of the models and sufficient amount of data for the given task. It can only be noted that normalization with dataset statistics results in greater smoothness and stability of the metrics curves for models trained from scratch, and the opposite effect for ResNet which was pre-trained on ImageNet. Examples of curves are shown in the Figure1.

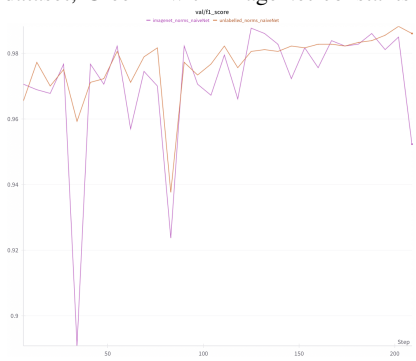
The results of the best experiments are summarized in TableI. From the results, it can be observed that pre-training on ImageNet is relevant, and that the task is solved with sufficiently high quality while utilizing only 13% of the original annotations.

Model	Tr. layers	Acc.	P	R	F1-score
VGG-like	All	0.980	0.989	0.989	0.989
C-ResNet	All	0.975	0.983	0.991	0.987
ResNet-18	L3,L4,FC	0.988	0.995	0.992	0.993

TABLE I: Results on the test set for the best models of each class selected on the validation dataset. Std calculated for different random seeds cannot fit in the table, however, they do not exceed 0.5% of the metrics. Tr. layers stands for trained layers, C-ResNet - Custom ResNet.



(a) F1-score on validation set for the custom ResNet trained from scratch: Purple - with constants gathered from the unlabeled dataset; Green - with ImageNet constants



(b) F1-score on validation set for the custom VGG-like model trained from scratch: Orange - with constants gathered from the unlabeled dataset; Pink - with ImageNet constants

Fig. 1: Normalization with dataset statistics results in greater smoothness and stability of the metrics curves for models trained from scratch, but doesn't show significant influence on the results

IV. Semi-supervised learning

To improve the results, it is possible to go further with the transfer learning approach and use larger models either CNN or ViT [4]. However, this approach is not of great research interest and will also increase the cost of inference for a fairly simple task. Thus, it is proposed to utilize 30.000 unlabeled pictures to improve the metrics. There are different approaches to working with such data. However, most of the self-supervised and unsupervised methods have been developed for object images and are likely to have limited adaptability to satellite images, where there is no clear distinction between foreground and background. Among unsupervised methods, the Autoencoders and Masked Autoencoders (MAE) [6] approach can be highlighted. This technique will allow to pre-train the ViT encoder-decoder (or in theory U-Net [12]) model on unlabeled data, and then fine-tune the pre-trained encoder on the annotated data, possibly obtaining thus an improved generalization. The main criticism of this train task in the literature is its complexity / redundancy to obtain discriminative features. There is also a question of the necessity and justification of using this method due to the high metrics of obtained supervised models.

Since the annotated part of the dataset allowed to obtain models with high metrics, the family of Active learning [1] [13] [15] and Semi-supervised learning (SSL) methods is of interest. The main idea of the methods is to use a pre-trained model to select two groups of images in an unlabeled dataset: informative/uncertain samples and high confidence samples. Uncertain samples are found based on different methods (entropy, Monte Carlo Dropout [5] and others) and then annotated for later inclusion in the training dataset. Since we cannot annotate the data in this work, we limit to the use of the second group. The model is confident in the classification since high confidence unlabeled samples are close to the labeled samples in the model's feature space. Therefore, the inclusion of these samples in the training set may not have as significant an impact as uncertain samples. However, the inclusion of pseudo-labeled images can play the role of data augmentation leading to robust features.

Various ways of using pseudo-labels in combination with supervised loss have been proposed in the literature. These include Pair loss [8] that implicitly propagates information between different unlabeled samples, the sample-relation-consistent mean-teacher framework [11], and the Unsupervised Data Augmentation (UDA) perturbation SSL method [16] [9]. In this work, the simplest method is used: for a pre-trained model, choose a confidence threshold value on the validation set, obtain pseudo-labels from the unlabeled dataset, add these cases to the training set, fine-tune the model and repeat.

The results of this approach are summarized in TableII. Confidence threshold is the minimum value by which the predicted probability must differ from 0.5 to assign a pseudo-label. Depending on the model and iteration of the algorithm, between 3.000 and 15.000 sampels were added to the training. Despite the improved results the plots reflect the theoretically predicted saturation, which is related to the proximity of the added images, such as in Figure2. It can also be of interest in future work to recalculate the Confidence threshold after each iteration of model training.

Model	Conf. th.	Δ Acc.	Δ f1-score	Δ FPs	Δ FNs
C-ResNet	0.4999	+0.8	+0.4	-63	+14
ResNet-18	0.499999	+0.3	+0.2	-1	-18

TABLE II: Metric difference (in percentage points for Accuracy and F1-score, and case counts for FPs and FNs) between ssl trained and supervised trained models on the test set. Custom ResNet and ImageNet ResNet-18 both pretrained on the train set and then trained with pseudo-labels chosen with the corresponding Confidence threshold.

V. Test-Time Augmentations and Ensembles

TTA [10] and ensembling [3] are classical ways to improve the accuracy and robustness of models at the expense of

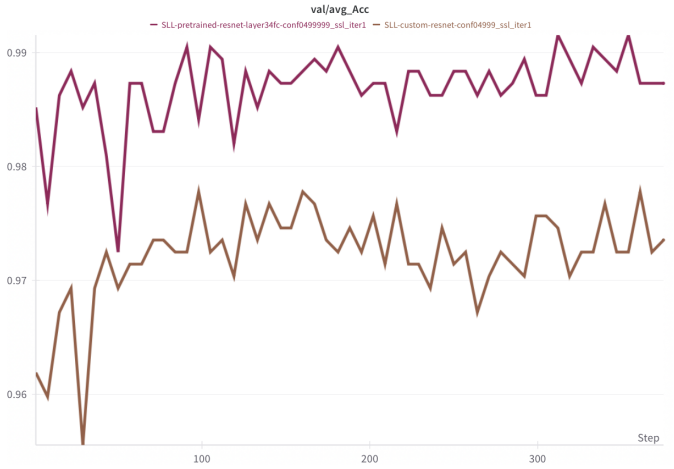


Fig. 2: Saturation effect of validation accuracy after 10 iterations of pseudo-labeling and fine-tuning of the pre-trained ResNet-18 (red) model and the pre-trained Custom ResNet (brown).

increasing the cost of inference. In this work, experiments were conducted with Majority Vote Deep Ensembles, which, however, did not show any improvement relative to the results of the best ensembled models. Perhaps a more careful selection of the set of models is needed. Test-Time Augmentations are less heavy than model ensembles, but it is still an effective method that applies data augmentation during testing to produce averaged outputs. However, the improvement from TTA depends on the model limitation of invariance to the used augmentations, also expected improvements decrease with the increase of the training dataset. [10] Thus, during the experiments, the typical improvement for trained from scratch supervised models was 0.02 percentage points (for f1-score for example) which corresponds to a decrease of 3-5 FP/FN cases. For pretrained ResNet-18 and SSL finetuned models, TTA did not improve the results.

REFERENCES

[1] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258, 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15099–15108, 2021.

[9] Tri Huynh, Aiden Nibali, and Zhen He. Semi-supervised learning for medical image classification using imbalanced training data. *Computer methods and programs in biomedicine*, 216:106628, 2022.

[10] Masanari Kimura. Understanding test-time augmentation. In *International Conference on Neural Information Processing*, pages 558–569. Springer, 2021.

[11] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging*, 39(11):3429–3440, 2020.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[13] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[16] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2(6):7, 2019.