Dual-stream Fusion and Data Augmentation for Imbalanced SAR and EO Multi-modal Classification

February 3, 2023

For this challenge, due to the large difference between SAR and EO images, we constructed a dual-stream network structure to encode the SAR and EO images seperately to feature vectors, and then concatenated the two vectors to an FC layer to generate the class prediction. We apply MobileNetV2 [1] structure with pretrained weights from timm [2] library as encoders for reasons from two aspects: 1) efficiency and lightweight, which can reduce the model size (compared with resnet34 based model (170.6M), MobileNetV2 based model is only 18.5M), and boost training speed while maintaining or even slightly boosting the performance. 2) reduce overfitting. Because larger models tend to overfit in this task dataset.

A main obstacle about this task is the imbalance of dataset of each class. To deal with this issue, we firstly use under-sampling to select same number of k samples for each class, k is set to the minimum of number of samples per class. Then 20% of each class is randomly separated out as validation set while the remaining are used for training. Focal loss [3] is used as the loss function instead of commonly used cross entropy loss to cast more attention on hard samples to tackle with overfitting. Label smoothing and various data augmentation are also used in calculating focal loss for the same reason.

Besides, we utilized semi-supervised strategy in training as well. After the accuracy in validation set is stable, the test set was inferenced with the trained model, and the results with max probability higher than threshold is added to the training dataset with the inferenced class as their pseudo label. Then the model continued to train some epoches using the combined train dataset. The model weights with best validation accuracy are used for the final inference to generate results of test dataset.

Even all the strategies are taken to reduce the bias of model to classes with more samples, the final outputs still tend to bias for the common classes (i.e. the first 4 classes which has over 10k samples) over long-tailed classes (i.e. last 6 classes, which all under 1k, except class box truck under 2k). Inspired by [4], we re-calibrated the output using post-processing: the first four classes are sorted by the predicted probabilities and labeled one-tenth of test samples for each class. Then the remaining samples which has relatively lower probabilities of the common classes are labeled by their maximum probability class. Experiments shown that removing the influence of common classes can obtain a more balanced result for the long-tailed classes.

Code is written in Python, using Pytorch framework. Experiments all done on desktop computer with ubuntu 18.04 system, Nvidia 3090 GPU with 24G memory, CUDA 11.5, CPU Intel(R) Core(TM) i9-11900K @3.50GHz. Training takes about 2 hours, and inference time of total 826 images is 8s (9.69ms/image pair). Implementation of all codes takes about 3 man-day. No model ensemble strategy is used.

References

[1] Sandler, Mark, et al. ”Mobilenetv2: Inverted residuals and linear bottlenecks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[2] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.11

[3] Lin, Tsung-Yi, et al. ”Focal loss for dense object detection.” Proceedings of the IEEE international conference on computer vision. 2017.

[4] Yang, Lehan, and Kele Xu. ”Cross Modality Knowledge Distillation for Multi-modal Aerial View Object Classification.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.