Fine grained Kaggle Challenge: Bird Classification

Nicolas DUFOUR
ENS Paris Saclay
nicolas.dufourn@gmail.com

Abstract

This data challenge aims to solve the problem of the classification of 20 species of birds. This task is particularly challenging because of the low data setup and the similarity between classes. This is also known as Fine-Grained Image classification.

1. Introduction

To do this we focus on 3 aspects. The first one is cropping the images around the birds. Then we are going to use a data augmentation pipeline to make the most of our dataset and avoid overfitting. Finally, we will use a classification model to predict the bird's species.

2. Dataset

The dataset is a subset of the CUB-200 dataset [1] where 20 classes are sampled. Exploring the dataset, we have seen 2 problems. First, the validation set is too clean: All the images are nicely framed, the birds are visible. On the contrary, the test set is quite out of distribution. A lot of birds are small in the image, there are occlusion, human hands. To try to bridge the distribution gap, we are going to crop the images extracting the birds, but it is not going to solve all the problems

3. Bird cropping

To crop the birds, we are going to use the Detectron2 [2] library that allow to use instance segmentation and detection models. We use 2 models: Mask-R-CNN [3] and RetinaNet [4]. We leverage the performances of the 2 models. For each picture, we pick the most confident prediction of a bird on the 2 models. (If there is multiple birds predicted we keep the most confident one.) Doing this, we manage to successfully crop all the images of the dataset. Then for the cropping, we expand the bounding box to fit a square. Indeed, the model take square entries and to avoid stretching, we prefer to crop in a square shape.

4. Data Augmentation

Data augmentation is key to avoid overfitting our training. We use 2 data augmentation pipelines: One with Albumentation [5] where we vary the contrast, brightness etc… The second approach is called AugMix [6]. Augmix help with a regularization of the training changing the loss. For both pipelines we used some rotation, flipping augmentation. We also added occlusions with random black squares to help predicting the more challenging cases where we have a partial bird. We also added a resizing constraint to help with the resolution discrepancy.

5. Model

For this competition we used the Vision Transformer (ViT) [7]. Even if we tried some CNN models, ViT outperformed all the models we tried. We used a pretrained model on ImageNet [8]. We manage to finetune it using our heavy Data Augmentation pipeline. The best results where obtained training the whole network (and not by freezing layers). The optimizer used was SGD with learning rate 0.001 and momentum 0.9. We used batch size 16 for Albumentation data augmentation and 4 for Augmix (due to material constraint).

6. Results

We obtained the following results

	Validation set	Public Test set
ViT + Album	0.92	0.88
ViT + AugMix	0.98	0.86

We see that Augmix has the best performances on the validation set. To make sure we did not overfitted the validation set we used 5- Fold Cross validation and obtained the same result. However, the performances are worst on the Public data. Therefore, we submit these 2 models to prevent having overfitted the public test set on our best performing model.

7. Conclusion

We are satisfied of the performances having the constraint of not using a better pretraining. However, we could maybe improve using ensemble models. We tried doing BYOL [9] self-supervised learning on NABirds [10], but the computational constraint did not enable this approach to be trained.

8. References

[1] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S.. The Caltech-UCSD Birds-200-2011 Dataset.

[2] Yuxin Wu, Alexander Kirillov, Francisco Massa and Wan-Yen Lo, & Ross Girshick. (2019). Detectron2. https://github.com/facebookresearch/detectron2

[3] He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN*2017 IEEE International Conference on Computer Vision (ICCV)*.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, & Piotr Dollár. (2017). Focal Loss for Dense Object Detection.

[5] Buslaev, A., Iglovikov, V., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. (2020). Albumentations: Fast and Flexible Image Augmentations *Information*, *11*(2).

[6] Hendrycks, D., Mu, N., Cubuk, E., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty *Proceedings of the International Conference on Learning Representations (ICLR)*.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale *arXiv preprint arXiv:2010.11929*.

[8] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, & Michal Valko. (2020). Bootstrap your own latent: A new approach to self-supervised Learning.

[10] Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., & Serge Belongie. (2015). Building a Bird Recognition App and Large Scale Dataset With Citizen Scientists: The Fine Print in Fine-Grained Dataset Collection.