Object recognition and computer vision, assignment 3: Birds Classification

Abstract

Fine tuning is a straightforward and usually simple task. Here, the proposed dataset was small (less than a thousand images) and mostly unbalance. This challenge was solved using state of the art Vision Transformers and most recent advances in deep embedding learning. The implementation allows the visualization of the learned embedding as well as the training and validation metrics. A custom architecture was implemented and the repository is available at `https://github.com/ArthurZucker/PytorchTemplate` *in the kaggle branch. The proposed method reached* `89.60` *on the testing set, but is more interesting as a different approach to solve few-shot learning tasks.*

1. Introduction

1.1. Data

The dataset used is a subset of the *Caltech-UCSD Birds-200-2011 bird dataset* [7] containing 1702 images. The split between hidden test set contains 517. As it is custom, the images were normalized and converted to tensor before being fed to the network. *AutoAugment* transformation were also applied. One method also used semantic segmentation on the images to extract only the pixels corresponding to birds, but it did not yield relevant results and was thus removed. Data augmentation was a must given the small amount of data available.

1.2. Fine tuning

For each of the architecture presented in the paper, the last fully connected layer was set to identity and a custom fully connected layer was added, in order to classify for the number of classes available in the proposed sub-dataset. Pretrained models on ImageNet were loaded, and no layers were freezed.

1.3. Visualization

Result and the various runs are available at https://wandb.ai/zart/Recvis-Kaggle/. The **wandb** [1] library was used to observe the training metrics, visualize the learned embeddings and sample transformed images. Links to the best models will be available in 4. Embeddings were obtained using pytorch *hooks* attached before the classifier in each model.

2. Related Work

Current state of the art classification methods on images use ViT (Vision Transformers), which are based on attention mechanism. Fine tuning on such model is a valid approach. Given the size of the dataset, few-shot learning approaches were explored. Advanced in this field allowed to move from constrastive learning using triplet loss [5] to the design of particular loss function which allow to learn a more discriminative embedding. *Arface* [2] and its variants (Subcenter Arcface etc) are considered state of the art and were implemented in this context.

3. Proposed methods

- Baseline : pretrained `resnet50` [4]. The baseline is a pretrained RESNET50 [cite] on ImageNet. The corresponding learned embedding are shown in 1 in the upper left corner. The best model training available at `wandb`
- Vision Transformer : The implementation from [3] was used. It loaded the pretrained architecture defined in [3].
- Contrastive `resnet50` and ViT. Using the *arcface* implementation from [6], the loss acts as the classifier, and the last fully connected layer is set to identity for both architectures.

4. Results

The ViT performs best for every metrics both at validation and testing time. The proposed *arcface* loss gives a better embedding, more centered and discriminative embedding.

	resnet50	C. Resnet50	ViT	C. ViT
Validation accuracy	92.31	88.54	94.00	94.00
test accuracy	80.60	76.12	**89.67**	83.87

Table 1. Training results

References

[1] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 1

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition, 2019. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is

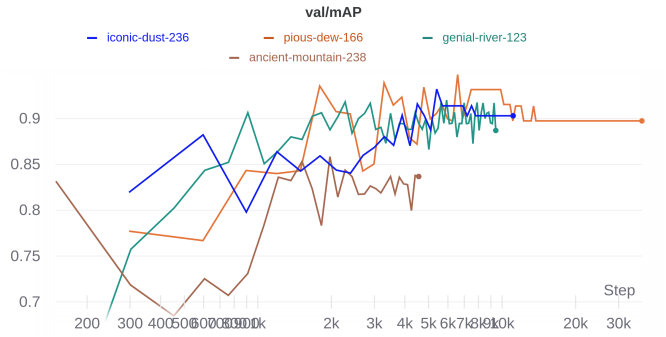

Figure 1. mAP results, best results are obtained with ViT implementations

worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[5] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network, 2018. 1

[6] Rong Huai, 2019. 1

[7] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1