Assignment 3 Report: Bird Classification Challenge

Eloi Tanguy
Ecole Polytechnique
eloi.tanguy@polytechnique.edu

Abstract

In this report, we present our strategy in model selection and training, maximising classification accuracy on a subset of the Caltech-UCSD Birds-200-2011 bird dataset.

1. Dataset

The provided data is separated into 1102 training images and 123 validation images all annotated into 20 bird classes. We naturally started training and validation following this division. However, given the severe class imbalance on the validation set and in order to control overfitting, we decided to separate all the data into 3 more balanced sets: 80% for training, 10% for validation, and finally 10% for testing.

Since we have very limited data, we also performed a variant of K-fold cross-validation for our model architectures: we performed trainings and validations on 10 randomly sampled train/val division. We experimented with voting models between the 10 trainings for set hyperparameters and enjoyed a slight (but costly) performance boost.

Furthermore, a peek at the unknown dataset for Kaggle revealed that the examples were more challenging, which brought us to perform geometric data augmentation on the training images. In general, we noticed that performing well with training augmentation equated to a better generalisation. Additionally, our models tended to struggle on certain classes (in particular the crows), thus we performed oversampling on the classes depending on our empirical estimation of their difficulty.

2. Model Selection

After a few preliminary baseline experiments (training simple models from scratch and fitting an XGBoost [1] classifier on CNN features), we quickly focused on fine-tuning models that were trained on ImageNet.

Our fine-tuning consisted in replacing the last linear layer in order to fit the 20 classes, which we trained alongside a variable amount of previous pre-trained layers. We tested the following models: AlexNet [6], VGG [8], ResNet(X) [4, 3], RegNet [11], EfficientNet [9], and Visual Transformers [10, 12]. Overall, better ImageNet accuracy led to better success with bird classification.

3. Training Parameters

We experimented with a wide variety of hyperparameters, starting with the learning rate and weight decay (searched along a logarithmic grid) and the dropout rate $p \in \{0, 0.1, 0.2, 0.5\}$. In general, the more a model could accept regularisation (weight decay and dropout), the better it performed on the test set.

We also experimented three different optimisers for the Binary Cross-Entropy loss: pure SGD, SGD with momentum $\mu = 0.9$ and Adam [5] (with default parameters). While Adam converged faster on the training set, models trained with SGD+momentum generalised better. Furthermore, we tried out several learning rate decay policies: constant, exponential decay $\gamma \in \{0.8, 0.9, 0.95\}$, cosine decay and cosine annealing with warm restarts [7].

We introduced random geometric transformations: rotation $|\theta_r| \leq \{20, 40, 80\}^\circ$ and shear $|\theta_s| \leq \{10, 30, 40\}^\circ$, translation of at most $\{10, 30, 50\}\%$ and scale $s \in \{[90, 110], [80, 130], [50, 150]\}$ % of the image.

4. Results

Our main successful models performed as follows:

model	Train	Val	Test	10-fold	Kaggle
ResNetX50	99.0	90.0	83.1	90.1	71.0
RegNet	96.9	92.4	89.1	91.3	76.1
ViT-L16	93.7	95.8	93.3	91.75	91.0

The Visual Transformer model is substantially larger, and was more robust to harsh data augmentation and regularisation, which explains its generalisation success on the more difficult Kaggle Test set.

One could do better by fine-tuning even larger models such as the convolution-attention hybrid CoAtNet [2] or larger versions of EfficientNet [9], which were far beyond our hardware capacity; or by training in stages with different levels of augmentation difficulty.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. 1

[2] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *CoRR*, abs/2106.04803, 2021. 1

[3] Wenfeng Feng, Xin Zhang, and Guangpeng Zhao. Resnetx: a more disordered and deeper network architecture. *CoRR*, abs/1912.12165, 2019. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 1

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[7] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. 1

[8] Karen Simonyan and Andrew Zisserman. 1

[9] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 1

[10] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677, 2020. 1

[11] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: Self-regulated network for image classification, 2021. 1

[12] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *CoRR*, abs/2106.04560, 2021. 1