Author response for “Fixing the train-test resolution discrepancy”

We thank the reviewers for their constructive feedback on the paper. We will take into account their comments on presentation and typos. Here we answer their main questions and comments.

R1: The results presented by the paper do not seem quite significant because just one dataset (ImageNet-2012) and two models (ResNet and PNASNet) are used. In addition, are the results shown significant? We will add experiments on additional models and datasets.

In particular, we have evaluated our approach for transfer learning for low-resource and/or fine-grained classification. Following common practice, (1) we initialize the model with weights obtained on ImageNet and (2) fine-tune it on the new dataset, which provides the baseline performance. Then (3) we use our method, i.e. we fine-tune the last batch norm and the fully connected layer with a higher resolution. Table 1 shows that for all transfer-learning tasks our method obtains a gain in accuracy of +0.4% to +1.3% absolute. All those results are obtained on competitive benchmarks where tens to hundreds of researchers have evaluated their methods, gains of around 1% accuracy are therefore significant.

Finally, we applied our method to a very large ResNeXt-101 32x48d from [Mahajan et al. ECCV’18], available online. We improved their top-1 Imagenet accuracy from 85.4% to 86.4%, which is the new state of the art on ImageNet.

R2: why not just fix the training such that there is no discrepancy, as opposed to changing the size for test and finetuning? We have tried other approaches to fix the discrepancy, which are not discussed for space reasons. The approach we presented consistently achieved the best performance while being simple and reducing the training time.

Regarding R2’s proposal, since the discrepancy is mainly due to the random resized crop data augmentation, we replaced it with a fixed size random crop and resize to K_{train}=224 pixel. When testing at K_{test} =224 we obtain 74.6% accuracy. The best test resolution is K_{test} =256: 75.2% accuracy, so there is an improvement at a slightly higher resolution. However, the result is significantly below the baseline accuracy obtained with random resize crop and without any resolution change (76.2%). This shows that it is important to keep a distinct data augmentation between train and test.

R2: Line 110-111 derives f = sqrt(HW), which does not seem to be right since k doesn’t include the sensor size. As often done for compactness (see [Hartley & Zisserman] eq. 5.9 or 6.9 depending on the edition), we expressed the focal length in unit of pixels; namely, the camera projection equation is $x = W^{\text{sensor}}_{\text{pixels}}/W^{\text{sensor}}_{\text{mm}} f_{\text{mm}} X/Z$ which simplifies to $x = fX/Z$ by setting $f = f_{\text{pixels}} = W^{\text{sensor}}_{\text{pixels}}/W^{\text{sensor}}_{\text{mm}} f_{\text{mm}}$. We will clarify this in the final version.

R3: what is a good ratio between the data augmentation hyperparameter(s) and the train/test size ratio? In other words, a general rule of thumb for practioners in image recognition on the ImageNet dataset. Thank you, we will add a “best practices” paragraph. It is hard to define a single ratio that works for all resolutions because the difference in statistics between the activations in the neural network does not vary linearly, eg. because of the padding of convolutional layers. An important effect to consider is the interaction between the input resolution and the size of the feature map before spatial pooling. There are key resolutions (largest size for a given feature map size) that provides local performance minima. For example, for a resnet-50 this occurs each time the resolution is a multiple of 32. Hence, a practical approach is to try out a few resolution steps above the training resolution, ie. $\{256, 288, .., 384\}$. This approach applies both with and without fine-tuning, but fine-tuning improves the effect.

R3: And, are the observations generalizable? If I transfer the ImageNet model to a specific task (e.g., CUB), will the findings in this paper be useful? If the answer is yes, how can it be useful? Yes, please see our answer to R1.

Dataset	Model	Baseline	Ours	State-Of-The-Art Models		
iNaturalist 2017	SENet-154	74.1	**75.4**	IncResNet-V2-SE	[Horn *et al.* ArXiv’17]	67.3
Stanford Cars	SENet-154	94.0	94.4	EfficientNet-B7	[Tan & Le, ArXiv’19]	**94.7**
CUB-200-2011	SENet-154	88.4	**88.7**	MPN-COV	[Peihua Li *et al.* ArXiV’19]	**88.7**
Oxford 102 Flowers	IncResNet-V2	95.0	95.7	EfficientNet-B7	[Tan & Le, ArXiv’19]	**98.8**
Oxford-IIIT Pets	SENet-154	94.6	94.8	AmoebaNet-B (6,512)	[Yanping Huang *et al.* ArXiv’18]	**95.9**
NABirds	SENet-154	88.3	**89.2**	PC-DenseNet-161	[Dubey *et al.* ArXiv’17]	82.8
Birdsnap	SENet-154	83.4	**84.3**	EfficientNet-B7	[Tan & Le, ArXiv’19]	**84.3**

Table 1: Transfer learning tasks. Evaluation is top-1, single crop and without changing the base Imagenet architecture.