Supplementary Material of "Tensor feature hallucination for few-shot classification"

Michalis Lazarou[1] Tania Stathaki[1] Yannis Avrithis[2]
[1]Imperial College London
[2]Athena RC

A. Datasets

***mini*ImageNet** This is a widely used few-shot image classification dataset [65, 50]. It contains 100 randomly sampled classes from ImageNet [30]. These 100 classes are split into 64 training (base) classes, 16 validation (novel) classes and 20 test (novel) classes. Each class contains 600 examples (images). We follow the commonly used split provided by [50].

CUB This is a fine-grained image classification dataset consisting of 200 classes, each corresponding to a bird species. We follow the split defined by [6, 19], with 100 training, 50 validation and 50 test classes.

CIFAR-FS This dataset is derived from CIFAR-100 [29], consisting of 100 classes with 600 examples per class. We follow the split provided by [6], with 64 training, 16 validation and 20 test classes.

When using ResNet-18 as a backbone network, images are resized to 224×224 for all datasets, similarly to other data augmentation methods [34, 8, 7, 40]. When using ResNet-12, they are resized to 84×84, similarly to [70].

RECONSTRUCTOR NETWORK	
Layer	Output shape
Input	$512 \times 7 \times 7$
ResBlockA	$256 \times 14 \times 14$
ResBlockA	$128 \times 28 \times 28$
ResBlockA	$64 \times 56 \times 56$
TranspConv3x3, stride=2	$64 \times 113 \times 113$
ResBlockB	$3 \times 226 \times 224$
Bilinear interpolation	$3 \times 224 \times 224$

Table 7. *Image reconstructor architecture*. ResBlockA is exactly the same as ResBlockB except that it uses ReLU activation function, while ResBlockB uses sigmoid.

B. Image reconstructors

We carried out an experiment to investigate whether the output tensor features without global average pooling (GAP) can provide more spatial information to aid the reconstruction of the original image, when compared to vector features obtained by GAP. A similar experiment has been carried out by [66] to visualize the tensor feature maps. We train two image reconstructors using a variant of an inverted ResNet-18 architecture with an additional transposed convolution layer, as shown in Table 7. The first is a tensor reconstructor, exactly as in Table 7. The second is a vector reconstructor taking a $512 \times 1 \times 1$ input. It is identical, except that it begins with an additional upsampling layer to adapt spatial resolution to 7×7.

We train each image reconstructor separately, taking as input the features as provided from the pre-trained ResNet-18 backbone, with and without GAP. For fair comparison, both reconstructors use exactly the same training settings, with ℓ_1 reconstruction loss as the loss function, batch size 128, Adam optimizer with an initial learning rate of 0.01 and 500 epochs with learning rate decreasing by a factor of 4 every 100 epochs. Similarly to Figure 1, is evident from Figure 5 that images reconstructed from tensor features are perceptually more similar to the original. The same holds for *generated* tensor and vector features, as shown in Figure 6. This experiment is for visualization purposes only; these images are not used in any way by our method.

Figure 5. *CUB images reconstructed from tensor/vector features of original images*. Each set of 3 rows depicts the original images (row 1), followed by the images reconstructed by the tensor (row 2) and the vector (row 3) reconstructor. Meant for visualization only.

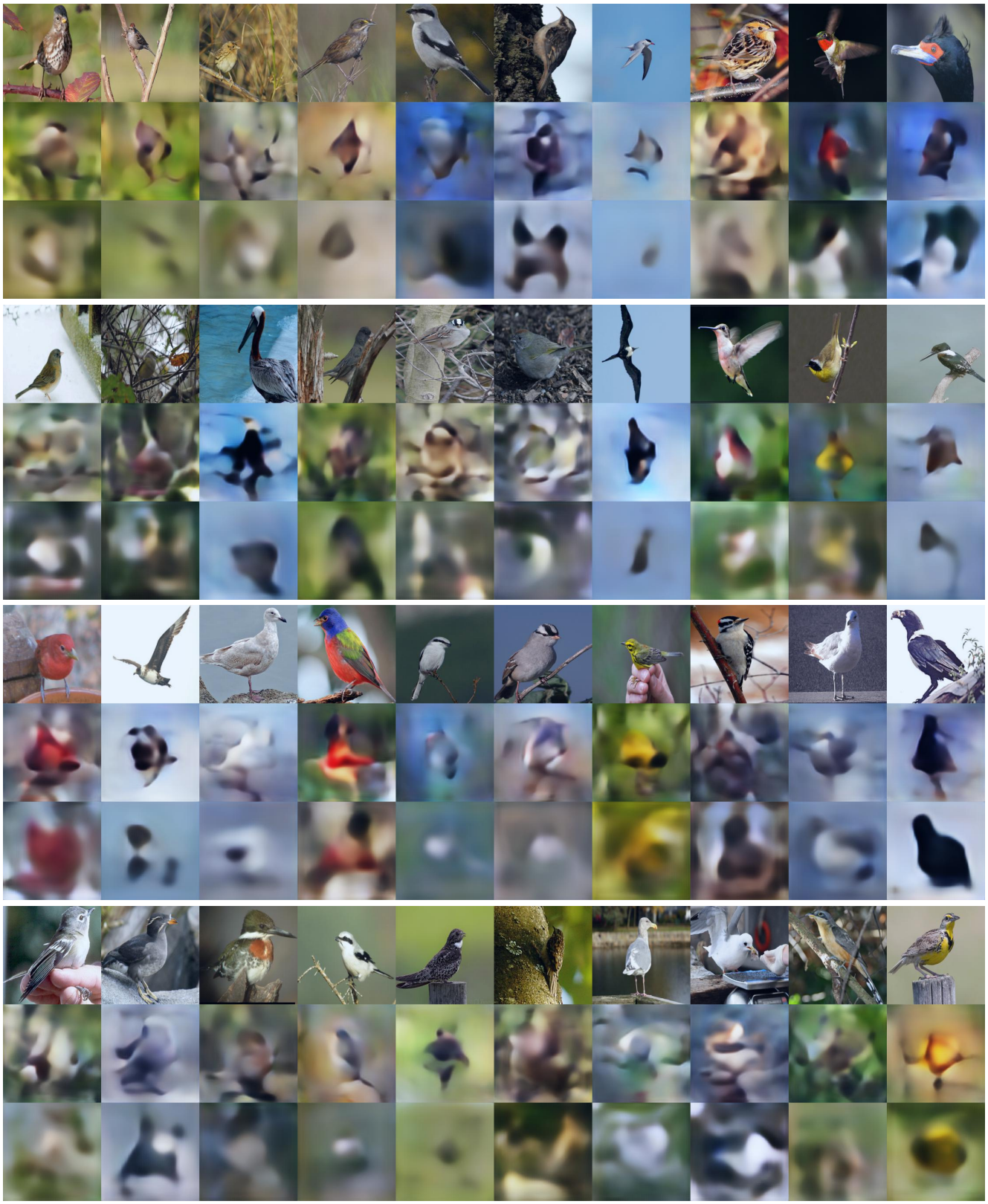

Figure 6. *CUB images reconstructed from our generated tensor/vector features*. Each set of 3 rows depicts the original images (row 1), followed by the images reconstructed by the tensor (row 2) and the vector (row 3) reconstructor. Meant for visualization only.