Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman
Visual Geometry Group, University of Oxford, UK

1. OVERVIEW

Given an image classification ConvNet, we aim to answer two questions:

- What does a class model look like?
- What makes an image belong to a class?

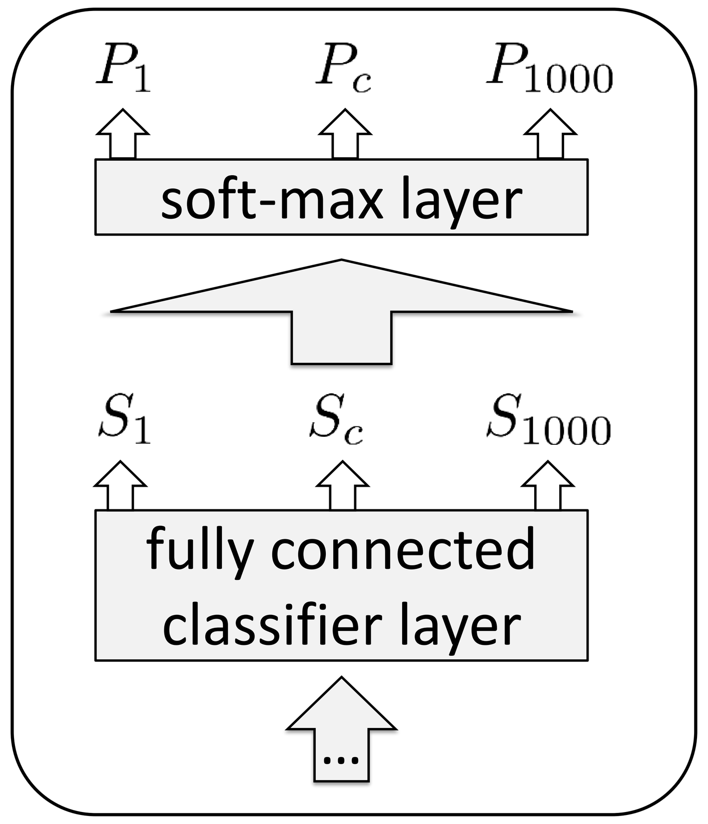
To this end, we visualise:

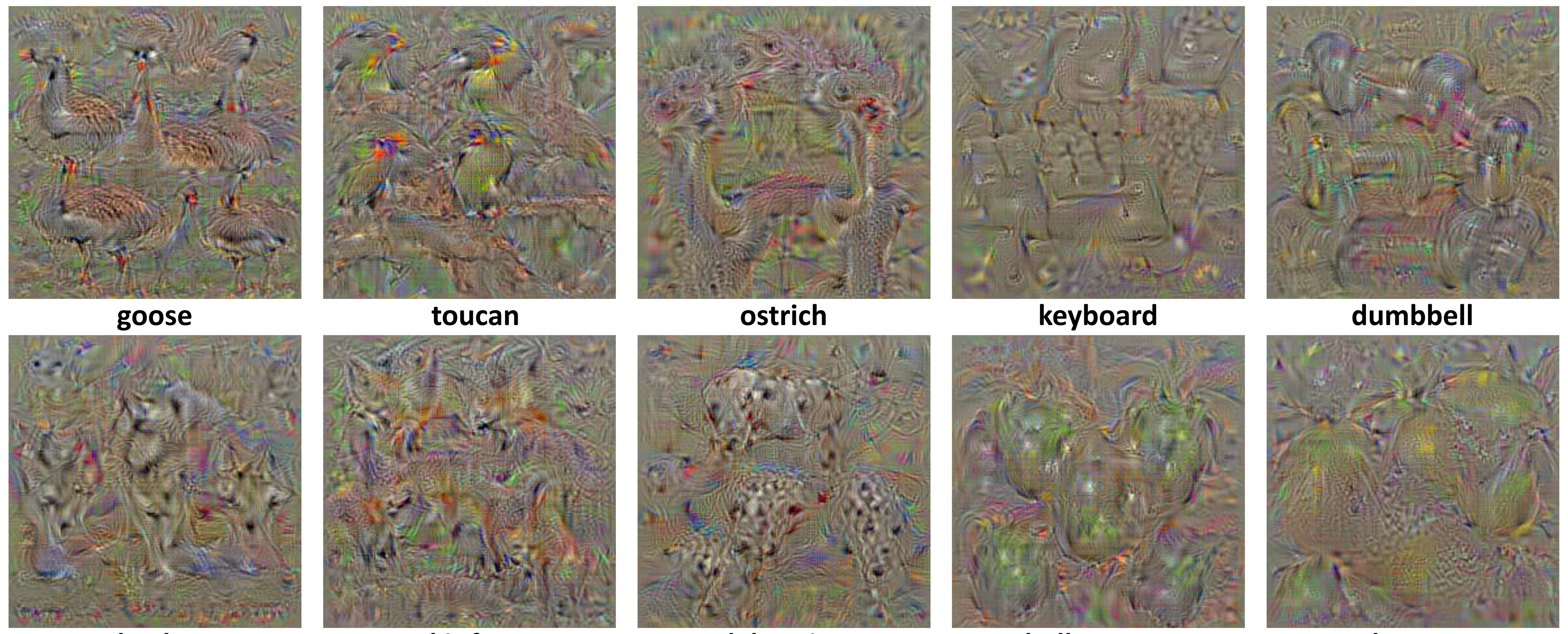
- Canonical image of a class
- Class saliency map for a given image and class

Both visualisations are based on the class score derivative w.r.t. the input image (computed using back-prop)

2. CLASS MODEL VISUALISATION

- We compute a (regularised) image I with a high class score $S_c(I) : \arg\max_I S_c(I) - \lambda \|I\|_2^2$ [Erhan et al., 2009]
- Optimised using gradient descent, initialised with the zero image
- Gradient $\partial S_c(I)/\partial I$ is computed using back-prop
- Maximising soft-max score $\arg\max_I P_c(I)$ leads to worse visualisation
- We visualise a ConvNet trained on ImageNet ILSVRC 2013 (1000 classes)





goose, toucan, ostrich, keyboard, dumbbell
husky, kit fox, dalmatian, bell pepper, lemon

3. IMAGE-SPECIFIC CLASS SALIENCY VISUALISATION

- Linear approximation of the class score in the neighbourhood of an image I_0 :

$S_c(I) \approx w^T I + b$ – score of c-th class

$$w = \left.\frac{\partial S_c(I)}{\partial I}\right|_{I_0}$$ – computed using back-prop

- w has the same size as the image I_0
- Magnitude of w defines a saliency map for image I_0 and class c

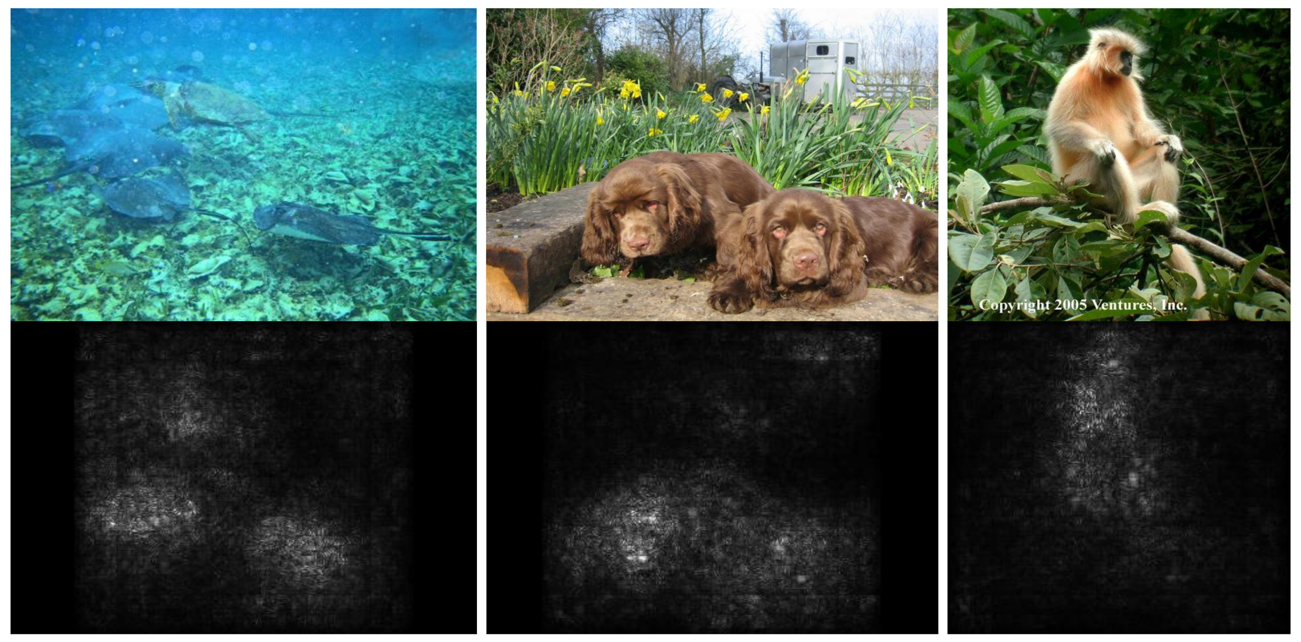
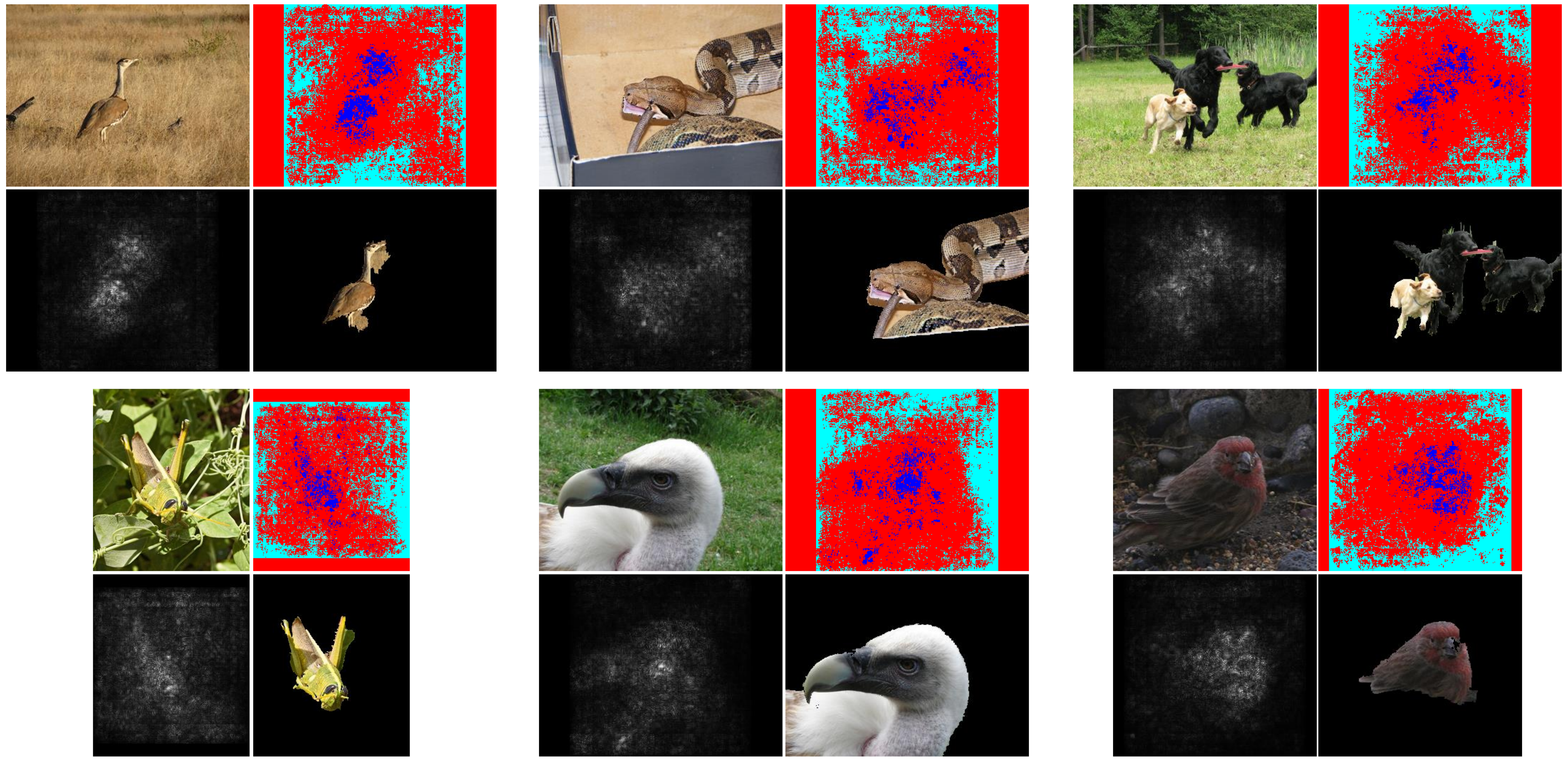


Image-Specific Class Saliency Properties:

- Weakly supervised
  - computed using classification ConvNet, trained on image labels
  - no additional annotation required (e.g. boxes or masks)
- Highlights discriminative object parts
- Instant computation – no sliding window, just a single back-prop pass
- Fires on several object instances

4. WEAKLY-SUPERVISED OBJECT LOCALISATION

- Given an image and a saliency map:
  1. Saliency map is thresholded to obtain foreground / background masks
  2. GraphCut colour segmentation [Boykov and Jolly, 2001] is initialised with the masks
  3. Object localisation: bounding box of the largest foreground connected component
- GraphCut propagates segmentation from the most salient areas of the object
- ILSVRC 2013 localisation accuracy: 46.4%
  - weak supervision: ground-truth bounding boxes were not used for training
  - saliency maps for top-5 predicted classes were used to compute five bounding box predictions



5. RELATION TO DECONVOLUTIONAL NETS

Layer	Forward pass	DeconvNet [Zeiler & Fergus, 2013]	Back-prop w.r.t. input
Convolution	$X_{n+1} = X_n \star K_n$	$R_n = R_{n+1} \star \widehat{K_n}$ equivalent	$\partial f/\partial X_n = \partial f/\partial X_{n+1} \star \widehat{K_n}$
RELU	$X_{n+1} = \max(X_n, 0)$	$R_n = R_{n+1}\mathbf{1}(R_{n+1} > 0)$ slightly different: threshold layer output vs input	$\partial f/\partial X_n = \partial f/\partial X_{n+1}\mathbf{1}(X_n > 0)$
Max-pooling	$X_{n+1}(p) = \max_{q \in \Omega(p)} X_n(q)$	$R_n(s) = R_{n+1}(p) \cdot \mathbf{1}(s = \arg\max_{q \in \Omega(p)} R_n(q))$ max location "switch" equivalent	$\partial f/\partial X_n(s) = \partial f/\partial X_{n+1}(p) \cdot \mathbf{1}(s = \arg\max_{q \in \Omega(p)} X_n(q))$

X_n – n_{th} layer activity; R_n – n_{th} layer DeconvNet reconstruction; f – visualised neuron activity

6. CONCLUSION

- Derivative of a ConvNet class score w.r.t. the input image is useful for visualising:
  - canonical image of a class
  - image-specific class saliency
- Image-specific class saliency can be further processed to perform weakly-supervised object segmentation and detection

This work was supported by ERC grant VisRec no. 228180
We acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU