Universal adversarial perturbations

Seyed-Mohsen Moosavi-Dezfooli*†
seyed.moosavi@epfl.ch

Alhussein Fawzi*†
alhussein.fawzi@epfl.ch

Omar Fawzi‡
omar.fawzi@ens-lyon.fr

Pascal Frossard†
pascal.frossard@epfl.ch

*The first two authors contributed equally to this work.
†École Polytechnique Fédérale de Lausanne, Switzerland
‡ENS de Lyon, LIP, UMR 5668 ENS Lyon - CNRS - UCBL - INRIA, Université de Lyon, France

A. Appendix

Fig. A.1 shows the original images corresponding to the experiment in Fig. 3. Fig. A.2 visualizes the graph showing relations between original and perturbed labels (see Section 3 for more details).

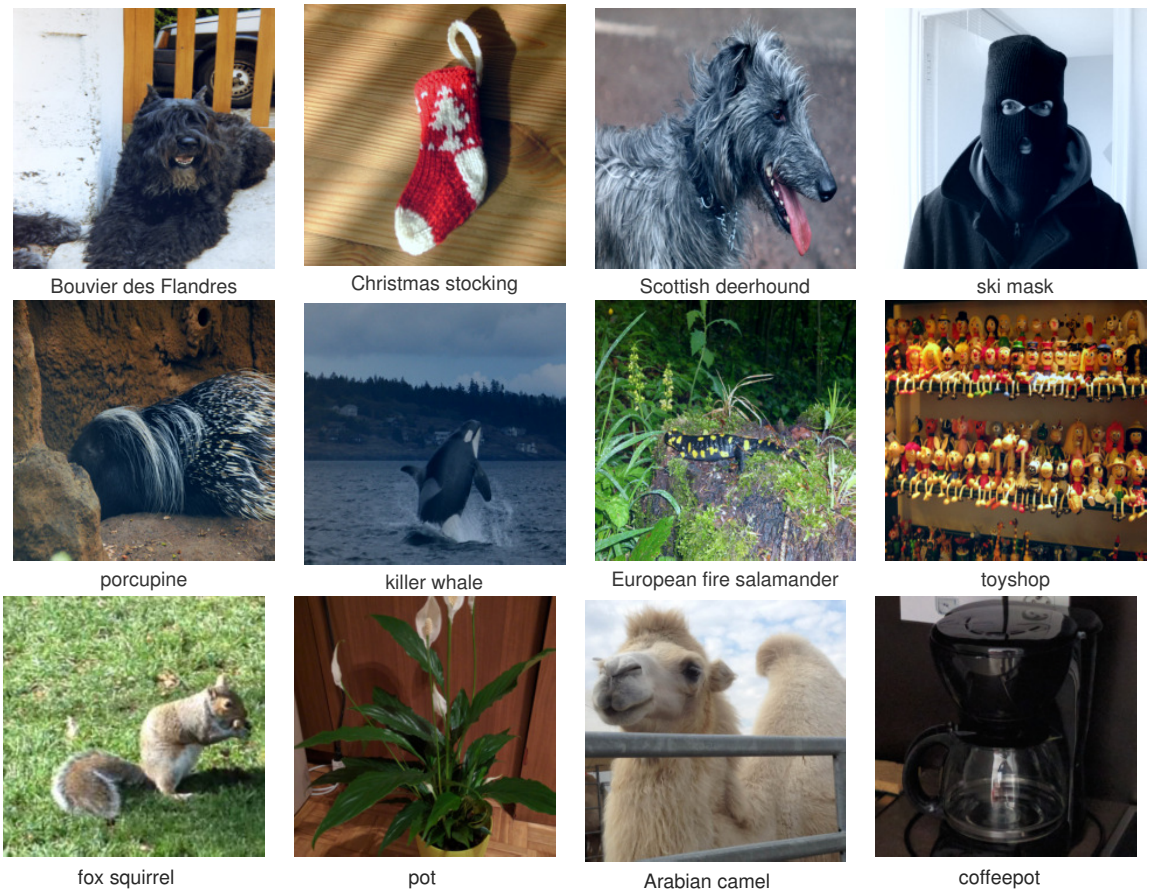

Figure A.1: Original images. The first two rows are randomly chosen images from the validation set, and the last row of images are personal images taken from a mobile phone camera.

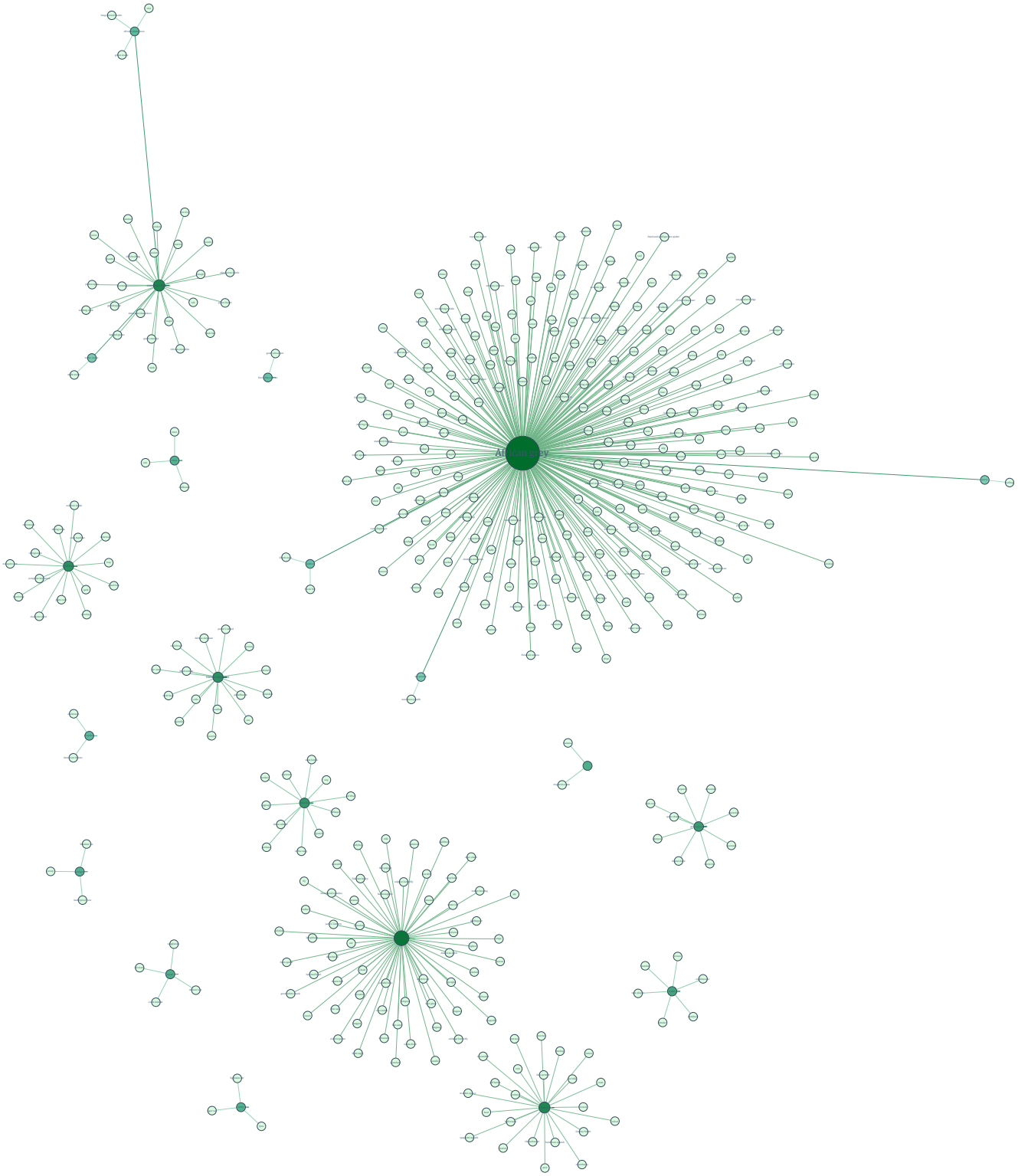

Figure A.2: Graph representing the relation between original and perturbed labels. Note that "dominant labels" appear systematically. Please zoom for readability. Isolated nodes are removed from this visualization for readability.