Pattern Recognition
and Applications Lab

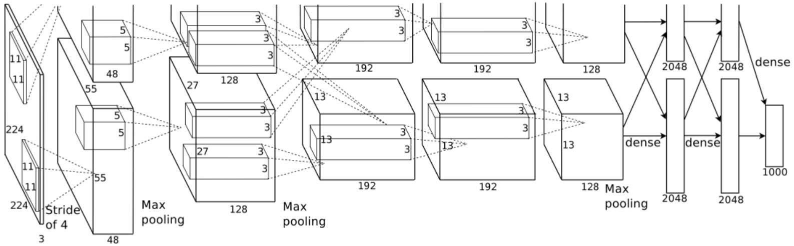
Adversarial Examples on ImageNet Classifiers

Battista Biggio

Department of Electrical and Electronic Engineering
University of Cagliari, Italy

ImageNet

- 1M images with 1,000 classes
  - 224x224x3
- AlexNet: DL winner in 2012

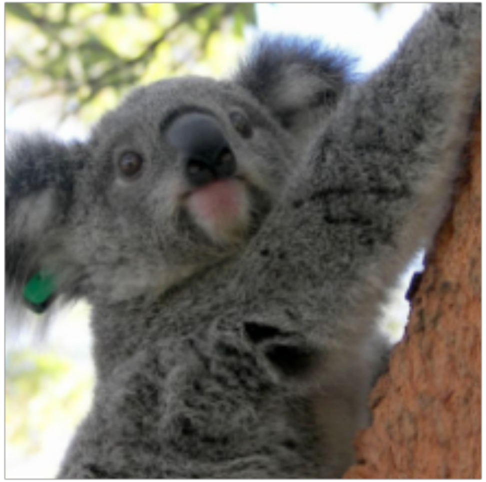


- Pretrained models available from torchvision

http://www.image-net.org/challenges/LSVRC/ (2012 edition)
https://en.wikipedia.org/wiki/ImageNet (historical remarks)
https://arxiv.org/pdf/1409.0575.pdf (ImageNet paper)
https://en.wikipedia.org/wiki/AlexNet (ILSVRC 2012 challenge winners with AlexNet)
https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (AlexNet paper)

Exercise 1: Pretrained ImageNet models

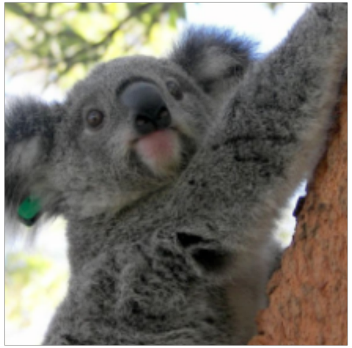
- Load a pretrained ImageNet model from torchvision, and classify an image
  - model = models.resnet18(pretrained=**True**)



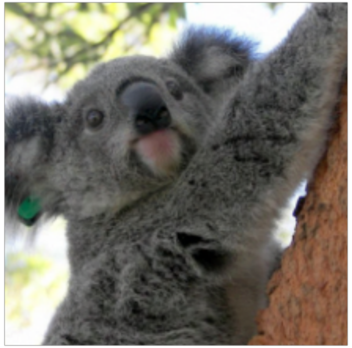
```
Label: koala               Score: 0.97
Label: ring-tailed lemur   Score: 0.02
Label: indri               Score: 0.01
Label: snow leopard        Score: 0.00
Label: titi                Score: 0.00
```

Exercise 2: Adversarial Examples

- Manipulate the *Koala* image to be misclassified as *acoustic guitar*
  - *see the notebook for details on the optimization process*



Label: **koala**	Score: **0.97**
Label: ring-tailed lemur	Score: 0.02
Label: indri	Score: 0.01
Label: snow leopard	Score: 0.00
Label: titi	Score: 0.00



Label: **acoustic guitar**	Score: **0.92**
Label: electric guitar	Score: 0.02
Label: banjo	Score: 0.01
Label: violin	Score: 0.00
Label: tabby cat	Score: 0.00