Common Objects in Context or Not

A dataset to test and analyse the failure modes of current Vision Algorithms

Max Wang, Mateus Zitelli, Stephen Gould.
Computer Vision has improved rapidly.

a couple of elephants standing next to each other on a beach
sand  beach  no person  seashore  sun  travel  sunset
ocean  desert  summer  island  fair weather  tropical
palm  vacation  hot
Recognition fails in circumstances trivial to humans
Recognition fails in circumstances trivial to humans
Recognition fails in circumstances trivial to humans
Recognition fails in circumstances trivial to humans

NeuralTalk captions a “red fire hydrant” for both
Recognition fails in circumstances trivial to humans

NeuralTalk detects giraffes in both images.
Good datasets fuel development

• Early datasets
  • Caltech 101 (2003)
    • 9,146 Images in 101 categories.
  • Caltech 256 (2006)
    • 30,607 Images in 256 categories.
  • Pascal VOC2009 (2009)
    • 14,743 Images
    • 20 classes.

• Modern datasets
  • Imagenet (2009+)
    • 14,197,122 images
    • 21841 synsets (sets of synonyms).
  • MS COCO (2014)
    • More than 300,000 images
    • with 80 objects in context annotated.
Current Challenges in CV – how does one measure performance?

- Classification, the identification of objects in images. Usually iconic and at a canonical perspective
Current Challenges in CV – how does one measure performance?

• Classification
• Localisation, drawing a bounding box over an iconic object
• Object Detection, drawing a bounding box over multiple objects
• Semantic Scene Labelling
• Captioning
Current Challenges in CV – how does one measure performance?

• Classification
• Localisation
• Object Detection
• Semantic Scene Labelling - Object Detection at a per-pixel scale.

University of Oxford

Objects appearing in the image:

- Bicycle
- Bird
- Dog
- Person
Current Challenges in CV – how does one measure performance?

• Classification
• Localisation
• Object Detection
• Semantic Scene Labelling
• Captioning – generating a description for an image

- a man is holding a stuffed animal in his mouth
- a couple of giraffes standing next to a tree
- a group of giraffes that are standing in the woods
- a woman holding a teddy bear in front of a mirror

"a cat is sitting on a couch with a remote control."
We should rethink the goals of Computer Vision

• What is important?
  • Robustness in classification over precision.
  • Localisation
  • Ability to understand and generalise concepts

• How can we achieve this?
  • New metrics for accuracy.
  • New challenges to test a framework’s understanding
  • A dataset which aids in analysing failure.
Our Proposed Solution - COCOONO

- We create a dataset consisting of permutations of a base image

Original

Clock removed

Airplane removed
Our Proposed Solution - COCOONO

• We create a dataset consisting of permutations of a base image

Original

Yellow alternative

Blue alternative
Our Proposed Solution - COCOONO

• We create a dataset consisting of permutations of a base image
Our Proposed Solution - COCOONO

• We create a dataset consisting of permutations of a base image
• This enables easy analysis of a vision framework behaviour on each set of images
• New “Challenges” for Computer Vision
COCOONO: Describe the Difference

The second image has a clock on the tower below the statue, however the plane in the clouds has been removed.
COCOONO: Describe the Difference

The snowboarder has been moved closer to the top of the snow covered slope from the bottom.
COCOONO: Real or Fake?

(-5.909907) a man riding a snowboard down a snow covered slope
Gathering the Dataset

• COCO took 70,000 person-hours.
• Crowdsourcing reduces cost, but gathering the dataset will still not be cheap/easy.
• We will automate as many tasks as possible.
Automating the Tasks

- COCO provided bounding polygons for objects
Image Inpainting

- Removes objects from the image
Image Reshuffling

- Moves objects around in an image
Image Recoloring

• Define classes which can be recolored and target colors
How we are currently gathering the dataset

• We setup a webapp where users evaluate automatically edited images.

• Technical details
  • Back-end: Node.js
  • Front-end: React.js w/ Redux
  • Database: RethinkDB
How we are currently gathering the dataset

• We integrate the Autodesk Pixlr online image editor which provides similar functionality to photoshop.
• Users can fix remaining inpainting problems.

• We would like to eventually develop an in-house image editing application with is more streamlined and possesses more useful tools.
Ensuring Dataset Quality

• Crowd-sourcers are unpredictable - ratings can be inconsistent. How do we ensure the quality of our dataset?
• Upon first login, users are shown a tutorial and ‘tested’
• Users are assigned a ‘reliability score’ which changes as they agree/disagree with other users
• Images are accepted only if the average grade is greater than 4.9 and the sum of the reliability scores of the graders exceeds some threshold.
Results so far

• 360,000 inpainted images.
• 2100 images evaluated.
• 600 good quality images (5 star or edited).
  And counting...
Results so far

Relation between inpainted area and grades

Grade

Square root of the inpainted area ($\sqrt{A}$)
Our Dataset on Neuraltalk

- Of 610 five star inpainted images, Neuraltalk captioned the removed object in 93 of them (15%)
Our Dataset on Neuraltalk

Most commonly mistaken inpainted objects:

- surfboard: 9 detected, 4 not detected
- skateboard: 5 detected, 3 not detected
- sink: 3 detected, 3 not detected
- bear: 7 detected, 2 not detected
- toilet: 8 detected, 3 not detected
- skis: 10 detected, 4 not detected
- clock: 3 detected, 4 not detected
- broccoli: 4 detected, 6 not detected
- giraffe: 9 detected, 10 not detected
- fire hydrant: 4 detected, 5 not detected
- kite: 10 detected, 6 not detected
- bird: 16 detected, 13 not detected
- airplane: 13 detected, 13 not detected
- frisbee: 3 detected, 10 not detected
- person: 10 detected, 19 not detected
- other: 415 detected, 97 not detected
Our Dataset on Neuraltalk

Normal Probability Plot for varea vs detecting removed object
Conclusion

• COCOONO aims to guide research along a path where true contextual understanding of an image is achieved.
Questions?


Desert Background from http://e2ua.com/WDF-943879.html

Neuraltalk: Karpnathy, A and Fei-Fei, L (2014), Deep visual-semantic alignments for generating image descriptions arxiv

Vision Frameworks Used:

Neuraltalk: https://github.com/karpathy/neuraltalk

Caffe: http://caffe.berkeleyvision.org/

Clarifai: https://www.clarifai.com/

Wolfram ImageIdentify: https://www.imageidentify.com/

University of Adelaide: http://demo.cs.adelaide.edu.au/

Universify of Oxford (CRFasRNN): http://www.robots.ox.ac.uk/~szheng/crfasrnn/demo/