Do you have a labeling problem?
Three tools for labeling data

ESIP Machine Learning Cluster
Labeling in AI

- Classification: the problem of identifying to which of a set of categories/features a new observation belongs

- A type of *supervised* learning

- Steps
  - Identify a set of samples from the data space, and tag them with one or more labels/categories
  - Train your model/algorithm
  - Provide new, unlabeled data to the model for it to predict the category
The labeling bottleneck

• Labeling is difficult to automate, often requires humans

• Many labeled samples may be needed, minimizing amount is hard and requires expertise

  • Imagenet, used in computer vision, the largest labeled dataset available to the public (http://image-net.org/update-sep-17-2019)

    • has 14,000,000 images

    • 22,000 visual categories

    • started with labels generated automatically from captions and tags

      • labeling took 50,000 paid workers looking at 160,000,000 images

• This need is holding AI back

• Tools, services are being developed
Presenters

• **Image Labeler**, Rahul Ramachandran, NASA

• **Labelimg**, Ziheng Sun, George Mason University

• **Bokeh**, Jim Bednar, Anaconda
Possible labeling tool, demos for Summer Meeting

- Arif, mathematical labeling approach
- Rahul, Image Labeler
- Katie, mobile-based tool
Extras
Labeling issues

• Knowledge of distribution of sample training data in the feature space is important, so that features encountered in training reflect real world distribution

• Human in the loop
  
  • a priori: training data corrected, validated by humans before training
  
  • ‘active learning’: select training data during training process to achieve specific accuracy,
    
    • does not solely rely on a priori, static assumptions
    
    • training process is more complex