DLI Accelerated Data Science Teaching Kit

Lecture 2.4 - Data Annotation and Data Quality

Data Annotation

Data annotation is the labelling of collected data for various applications.

- Image Processing
- Natural Language Processing
- Smart Grid
- … …

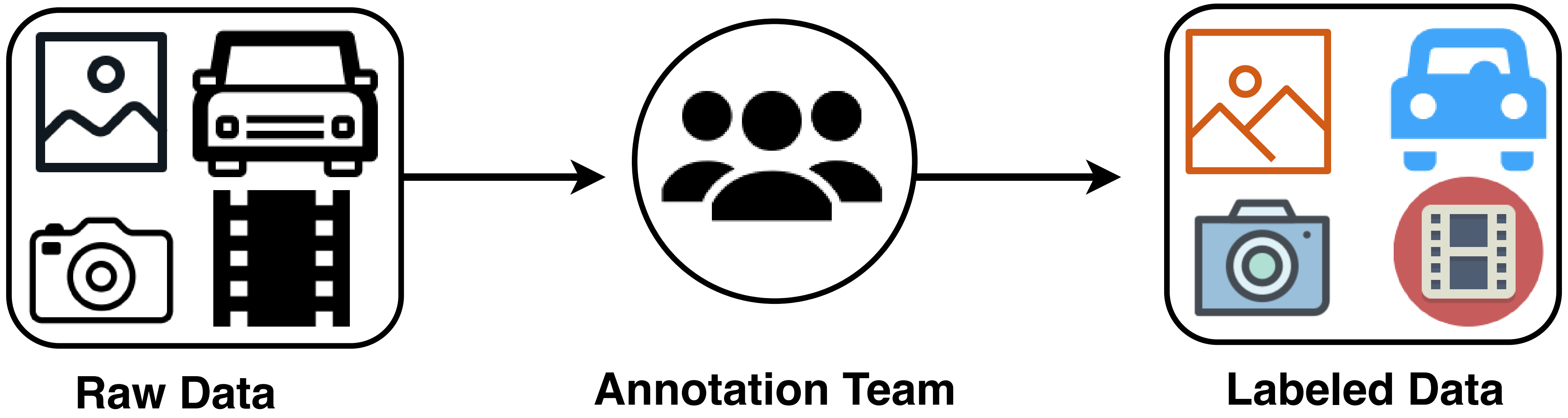
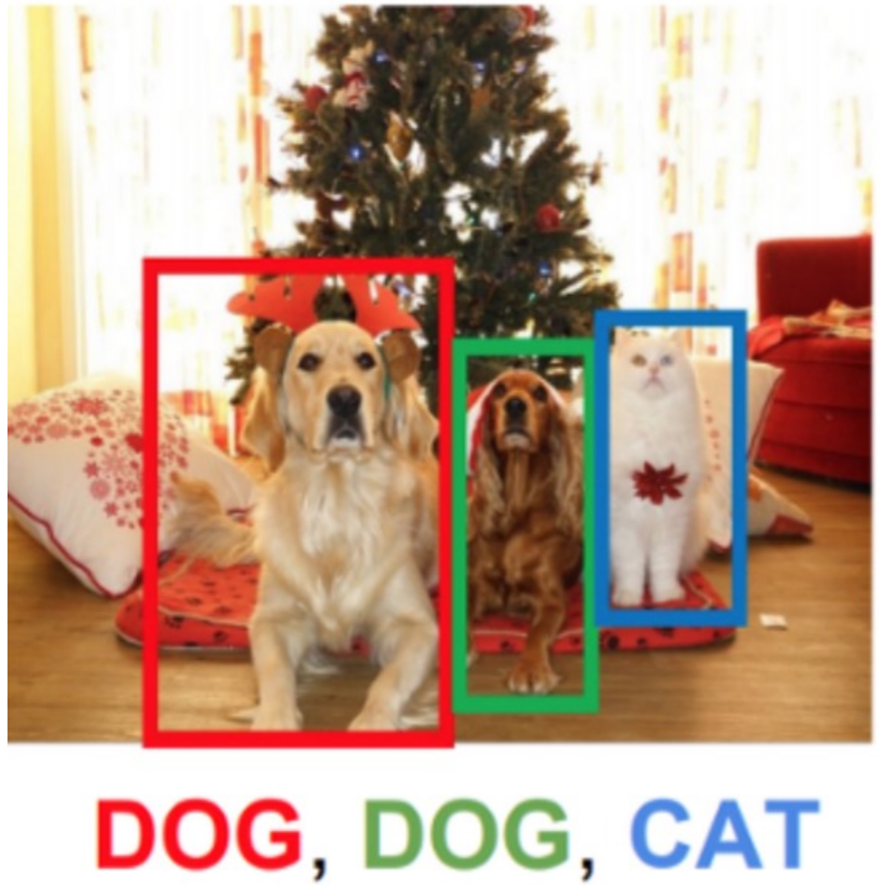


Image Processing

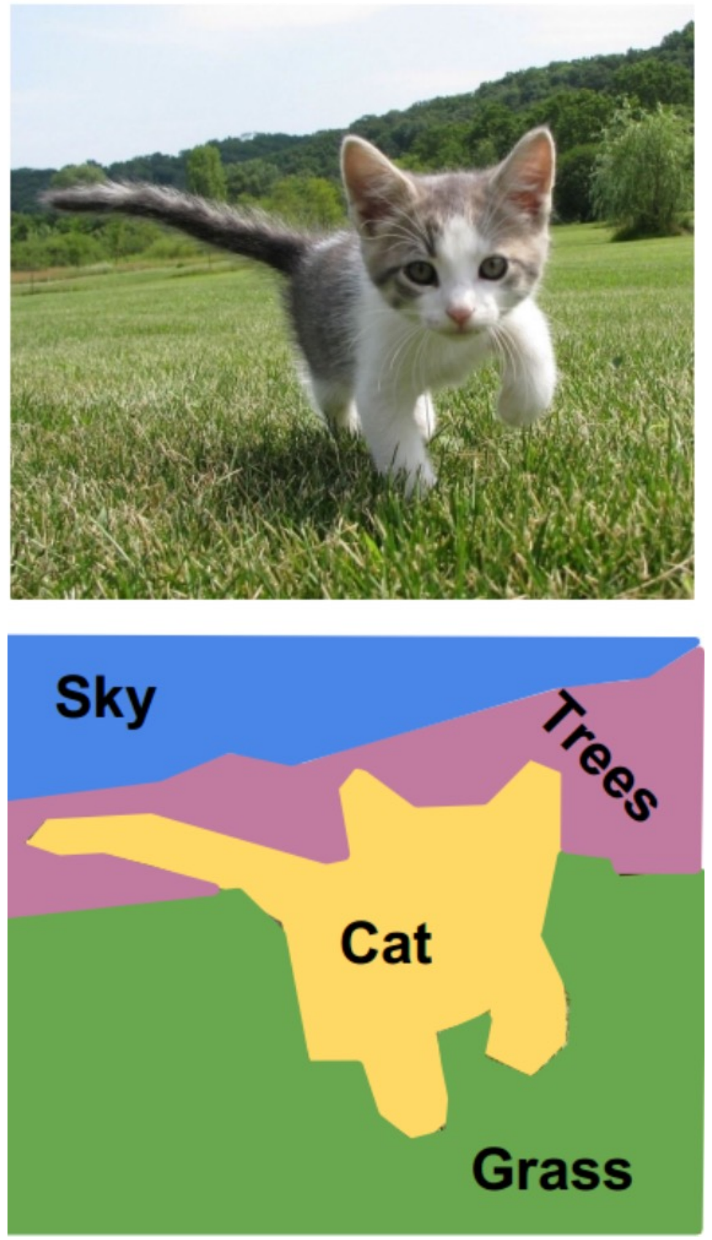
Examples of data annotation

Image Classification

Object Detection



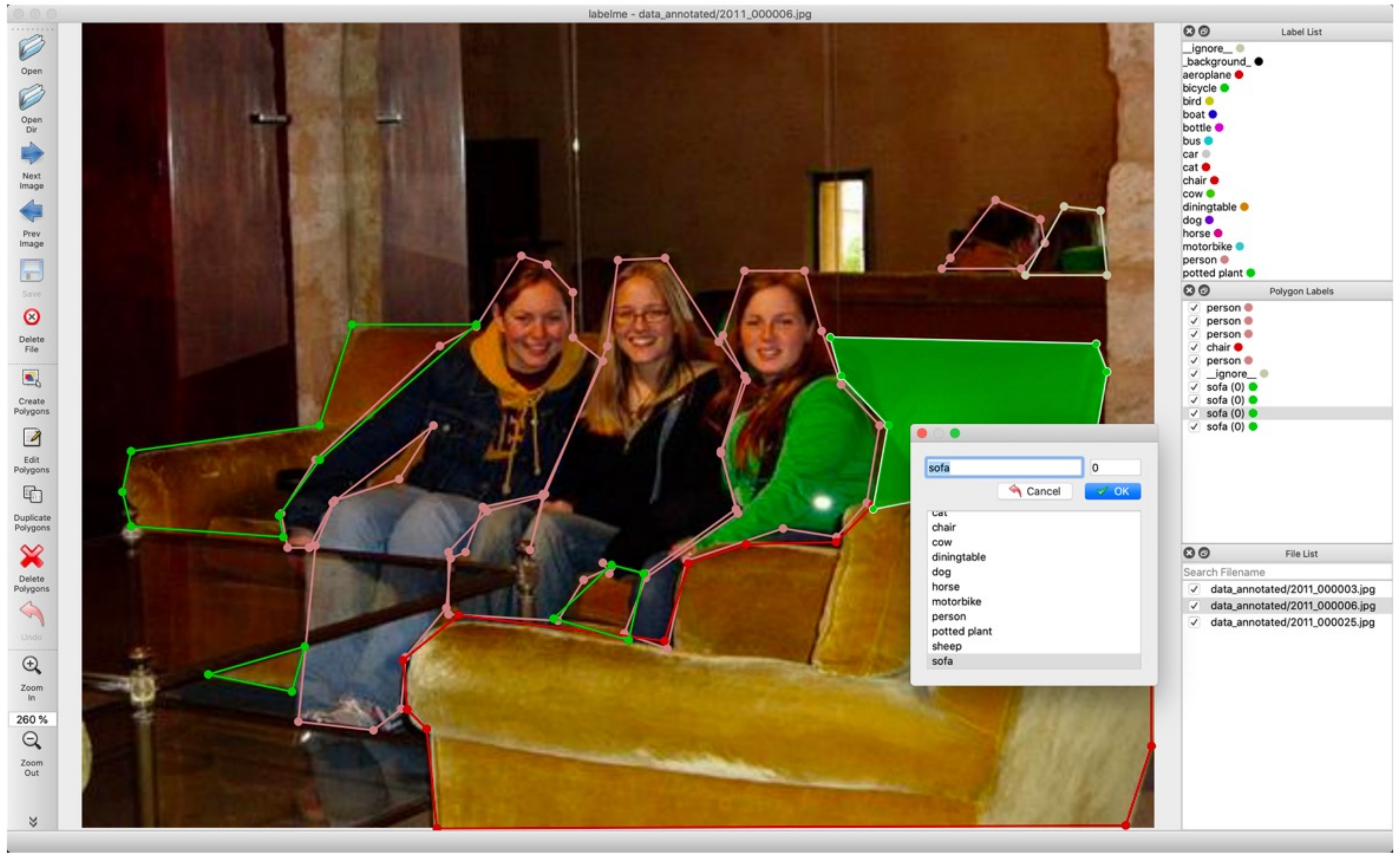
Semantic Segmentation



LabelMe

A graphical image annotation tool inspired by http://labelme.csail.mit.edu.

It is written in Python and uses Qt for its graphical interface.



Source: https://github.com/wkentaro/labelme

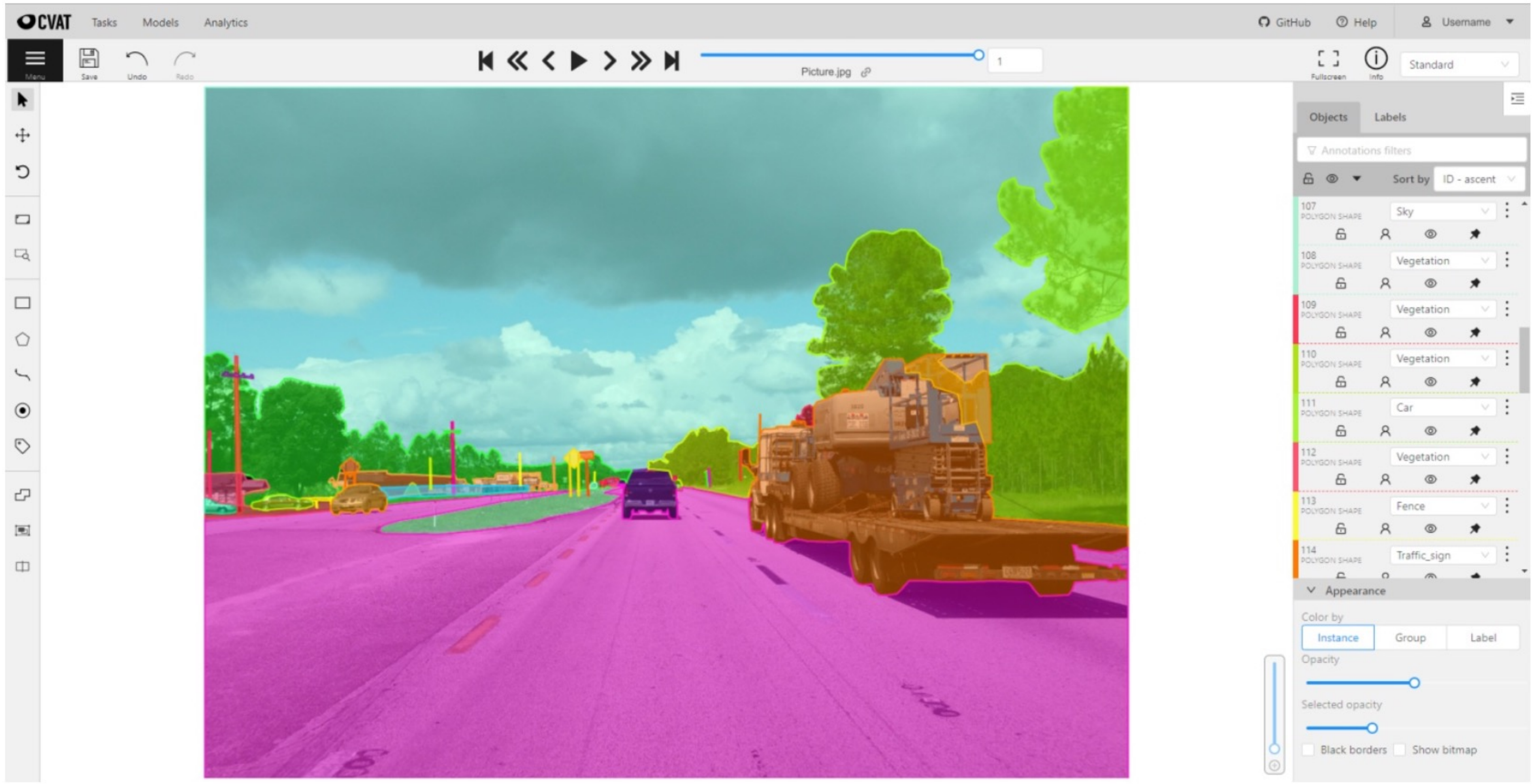
Computer Vision Annotation Tool (CVAT)

Open source

Auto annotation using trained DL models

Collaborative

Easy to deploy and maintain



Source: https://github.com/openvinotoolkit/cvat

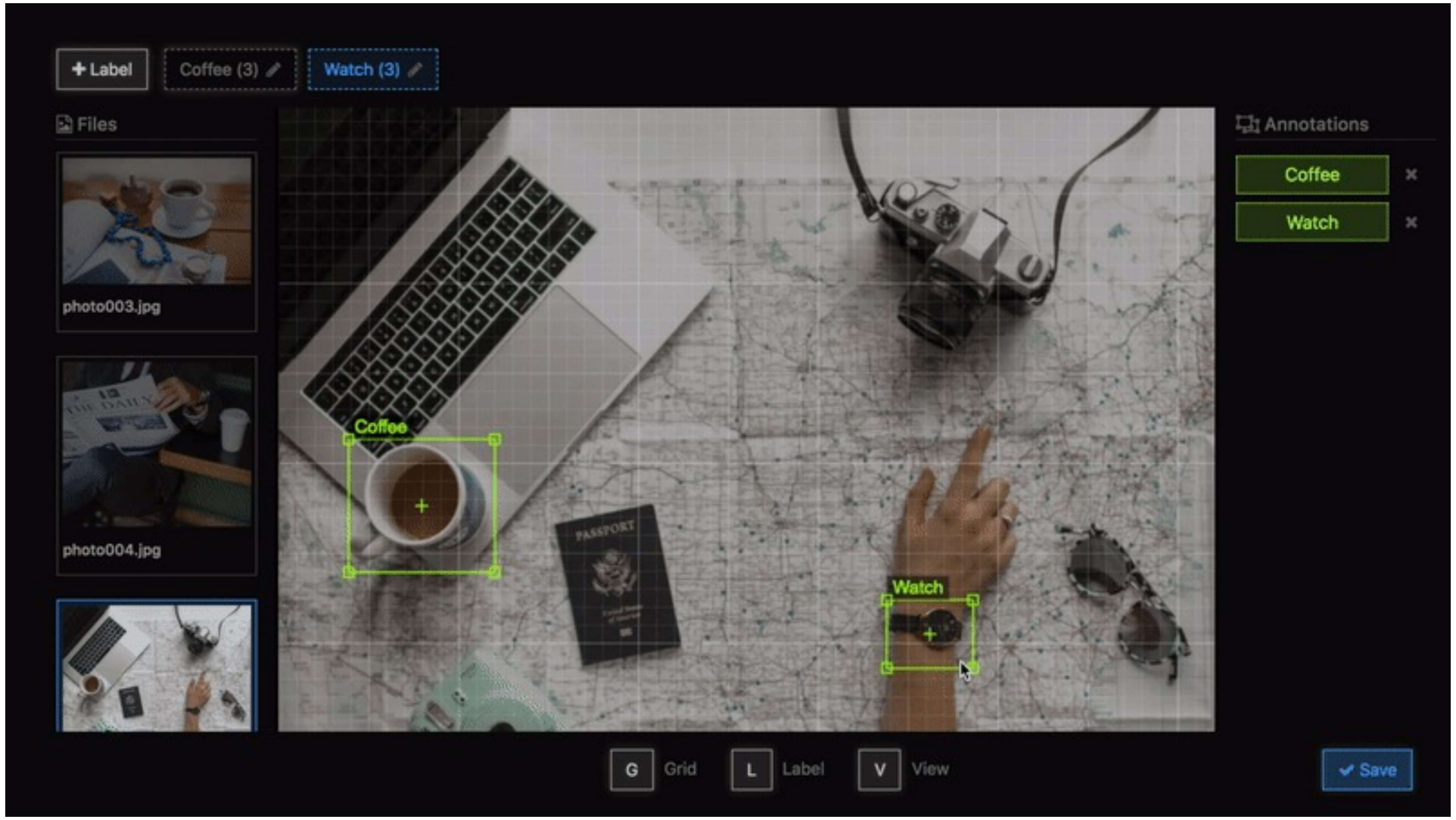
Einstein Vision Object Detection

Open Source

Image Size Change and Move

Label Change and Remove

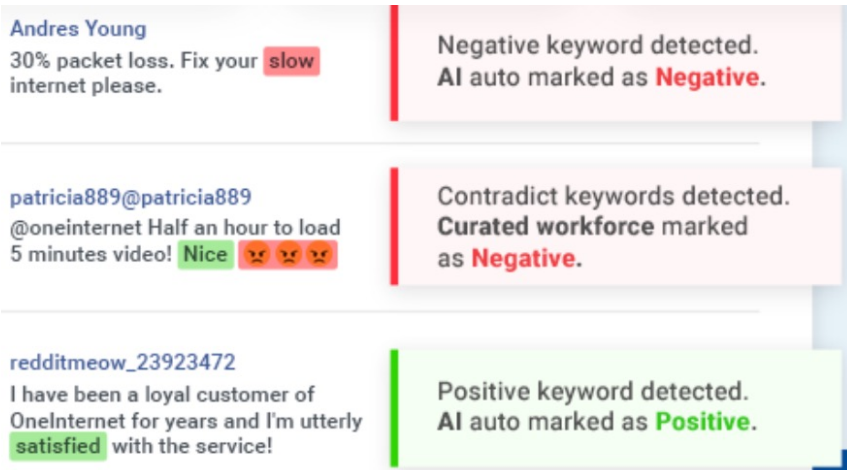
Easy to Use



Source: https://github.com/misu007/einstein-vision-object-detection-annotation-tool-sample

Natural Language Processing (NLP)

Sentiment Classification on Texts



Andres Young
30% packet loss. Fix your slow internet please.

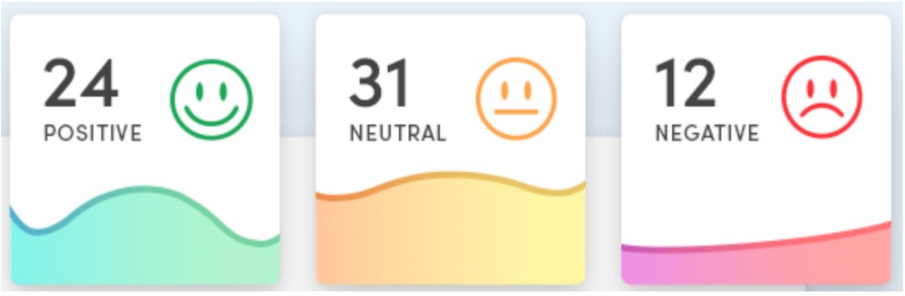
Negative keyword detected.
AI auto marked as Negative.

patricia889@patricia889
@oneinternet Half an hour to load 5 minutes video! Nice 😡😡😡

Contradict keywords detected.
Curated workforce marked as Negative.

redditmeow_23923472
I have been a loyal customer of OneInternet for years and I'm utterly satisfied with the service!

Positive keyword detected.
AI auto marked as Positive.

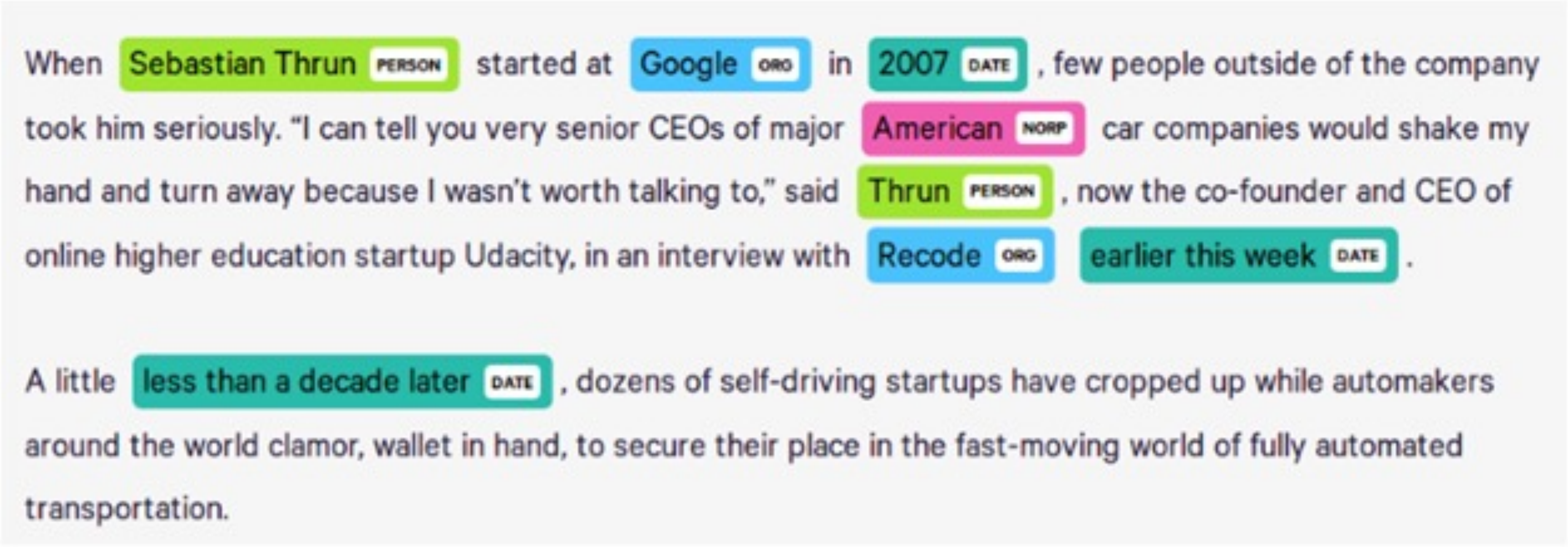


24 POSITIVE

31 NEUTRAL

12 NEGATIVE

Named Entity Recognition



When Sebastian Thrun PERSON started at Google ORG in 2007 DATE , few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Quality Control for Data Annotation

Calculating the accuracy of an annotator compared to ground-truth data.

Calculating the overall agreement and reliability of a dataset.

Measuring inter-annotator agreement on a per-task basis to generate a confidence score for each training data label.

Designing architectures that incorporate subject matter experts into the annotation workflow.

Breaking up a task into simpler subtasks to improve accuracy, efficiency and quality control.

Source: https://livebook.manning.com/book/human-in-the-loop-machine-learning/chapter-8/v-8/1

Data Quality (DQ)

Data quality refers to the state of qualitative or quantitative pieces of information.

Data is of high quality if it is “Fit for Use” in their intended operational, decision-making and other roles.

Quality Properties

- Relevance – does data meet basic needs?
- Accuracy – are key data elements correct?
- Timeliness – is data current?

Source: https://www.researchgate.net/profile/William_Winkler/publication/220695391_Data_Quality_and_Record_Linkage/links/551be3a90cf20d5fbde21e93.pdf

Data Quality (DQ)

Data quality refers to the state of qualitative or quantitative pieces of information.

Data is of high quality if it is “Fit for Use” in their intended operational, decision-making and other roles.

Quality Properties

- Comparability
  - Can several databases be combined into a data warehouse?
- Completeness
  - No missing records, no missing data elements

Source: https://www.researchgate.net/profile/William_Winkler/publication/220695391_Data_Quality_and_Record_Linkage/links/551be3a90cf20d5fbde21e93.pdf

Why Care About Data Quality?

High-quality data can be a major business asset, a unique source of competitive advantage.

Poor data quality can lower customer satisfaction.

Poor data quality can lower employee job satisfaction too, leading to excessive turnover and the resulting loss of key process knowledge.

Poor data quality can also breed organizational mistrust and make it hard to mount efforts that lead to needed improvements.

Source: https://www.researchgate.net/profile/William_Winkler/publication/220695391_Data_Quality_and_Record_Linkage/links/551be3a90cf20d5fbde21e93.pdf

An Example

Insurance companies need accurate and complete databases in order to:

- Accomplish Mandated Accounting Tasks
- Compute Insurance-in-force
- Compute Claim Rates
- Calculate Premiums
- Construct Models to Predict Future Results

Source: https://redrockslocksmith.com/do-insurance-companies-cover-car-lockouts/

How Do You Obtain High-Quality Data?

Prevention: Keep Bad Data Out of the Database/List

Detection: Proactively Look for Bad Data Already Entered

Repair: Let the Bad Data Find You and Then Fix Things

Allocating Resources: How Much for Prevention, Detection, and Repair

Source: http://content.schweitzer-online.de/static/catalog_manager/live/media_files/representation/zd_std_orig__zd_schw_orig/001/813/790/9780387695020_content_pdf_1.pdf

DEEP LEARNING INSTITUTE

DLI Accelerated Data Science Teaching Kit

Thank You