Can You See What I See? The Power of Deep Learning

Scott Thibault, Ph.D.
StreamLogic

Presentation Agenda

Three major vision tasks solved using deep learning:

- Image Classification
  - Answering questions about an image as a whole
- Object Detection
  - Locating objects within an image
- Embeddings
  - Measuring semantic similarity
- Conclusions

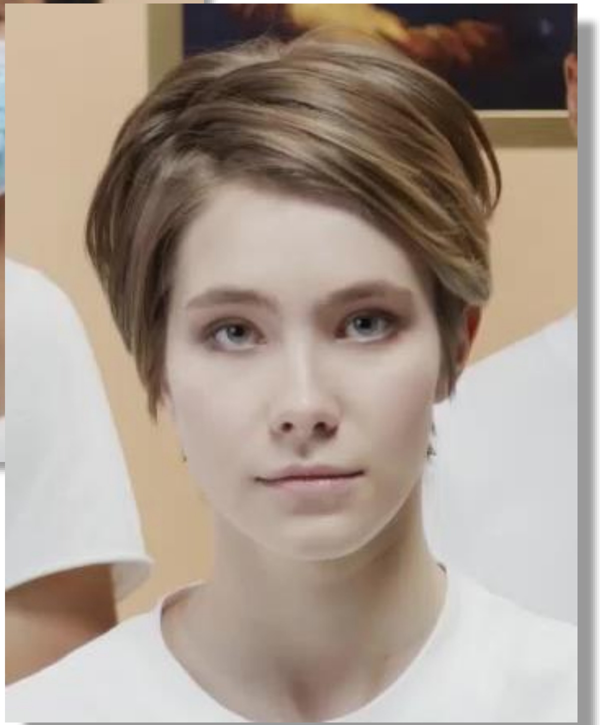
Image Classification

Image Classification

Answers the question for the whole image:

1. Binary classification – Is this image *X* (yes/no)?
2. Multi-class classification – Is this image *X*, *Y*, or *Z*?
3. Multi-label classification – What labels X, Y, and Z describe this image?

Mask (0.86)



No Mask (0.06)

Multi-Class Classification

Age classification output

Class	Probability
0-2	0.000011
4-6	0.000008
8-13	0.000187
15-20	0.001444
25-32	0.313656
38-43	0.683580
48-53	0.001012
60+	0.000101

streamlogic

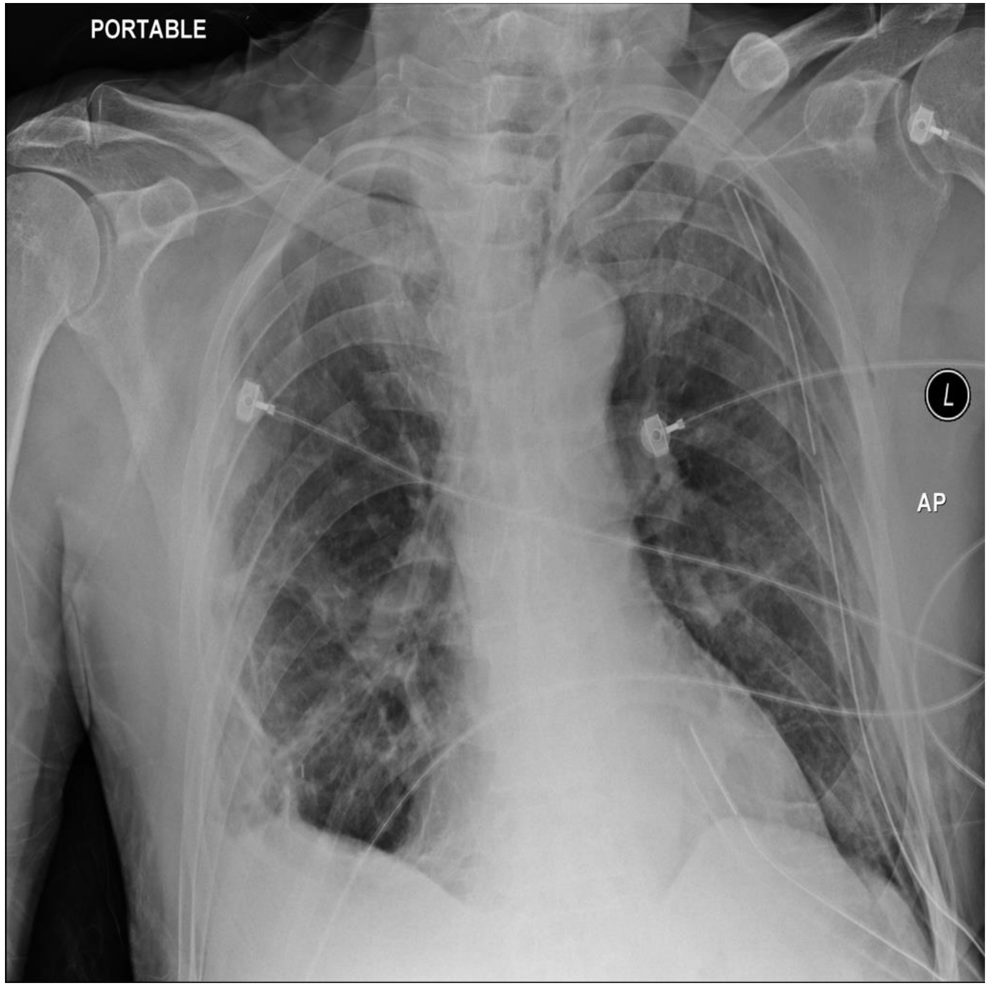
Multi-Label Classification

Multi-label classification output

Label	Probability
Smooth	0.00021
Rough	0.01332
Furry	0.87412
White	0.00340
Black	0.91021
Yellow	0.00013
Green	0.00002
...	...

Applications of Image Classification

- Medical diagnosis
- Marketing profiles
- Smart camera event capture
- Quality control
- Policy verification



Pre-Trained Image Classification Models

Public datasets with pre-trained models:

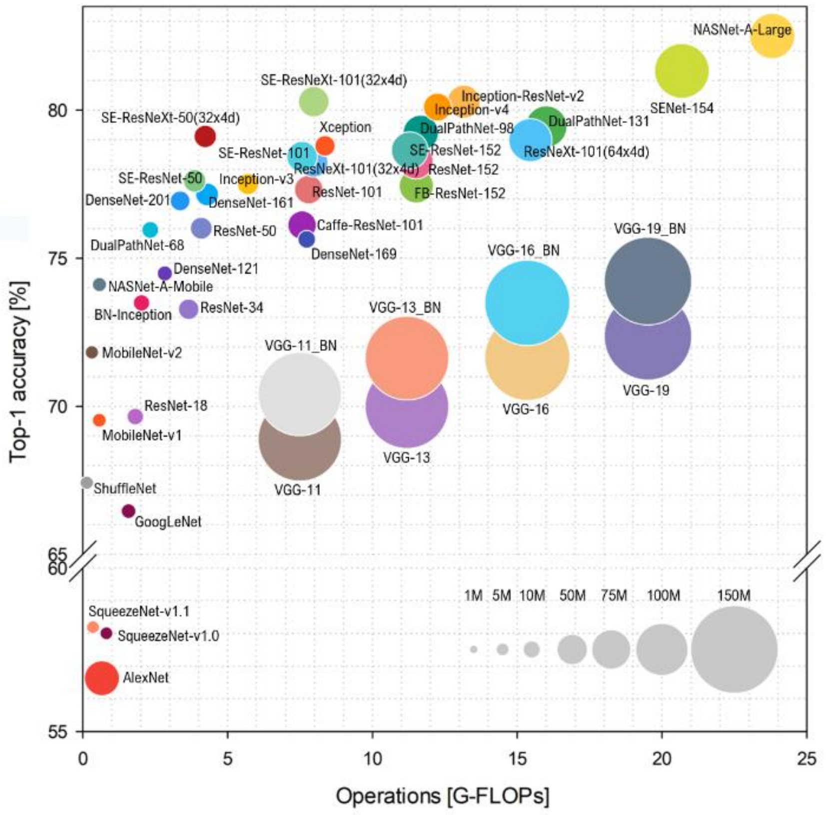
- ImageNet (ILSVRC) – 1000 classes
- COCO – 80 classes
- Pascal VOC – 20 classes
- Stanford Cars
- Places365
- Demographics (gender/age/race)
- Emotions (facial expression)
- Activities, e.g. sports
- Plants

Places365

Training Data for Image Classification

- Binary classification
  - For each image: one binary value (yes=1, no=0)
- Multi-class classification (“One-hot” encoding)
  - For each image: binary vector of length N with exactly one 1
- Multi-label classification
  - For each image: binary vector of length N with any number of 1s

Image Classification Accuracy



S. Bianco, R. Cadene, L. Celona and P. Napoletano, "Benchmark Analysis of Representative Deep Neural Network Architectures," in *IEEE Access*, vol. 6, pp. 64270-64277, 2018

How do we compare CNN architectures?

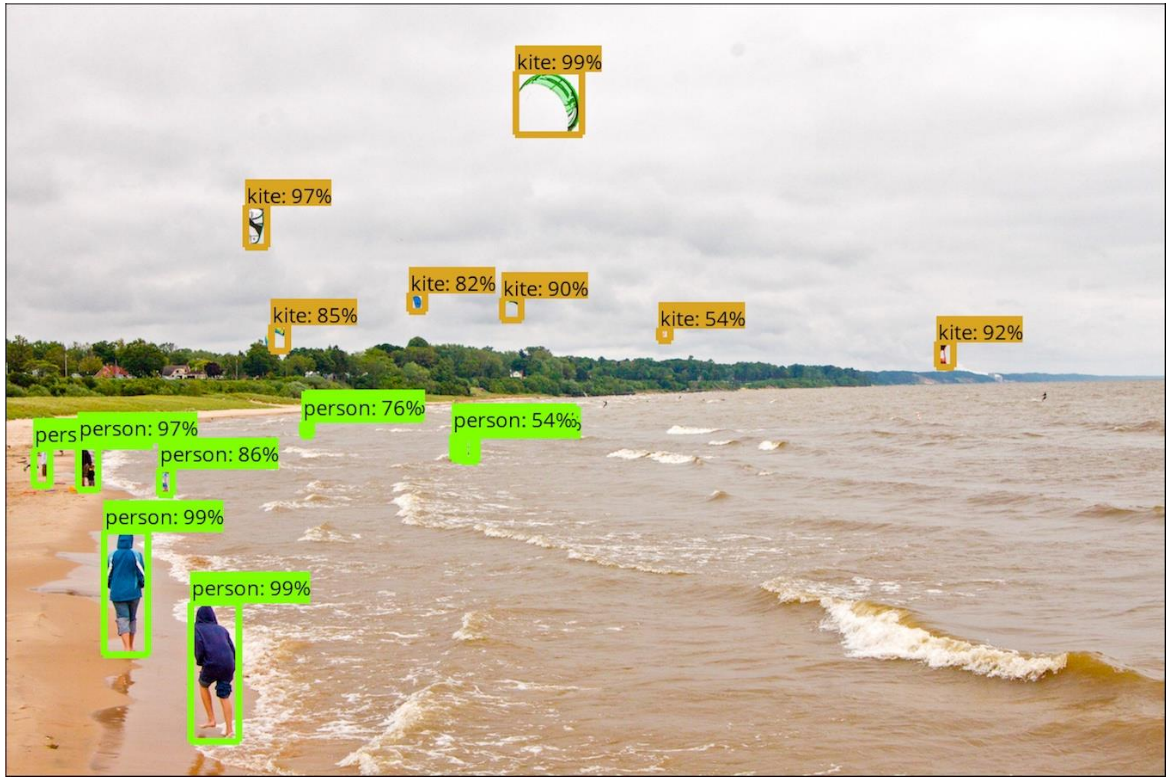
- ImageNet Benchmark
  - > 1M training images
  - 1000 classes
- Top-*N* Accuracy – correct class is in *N* highest class probabilities
- Model size affects accuracy
- Model size affects speed

Object Detection

streamlogic

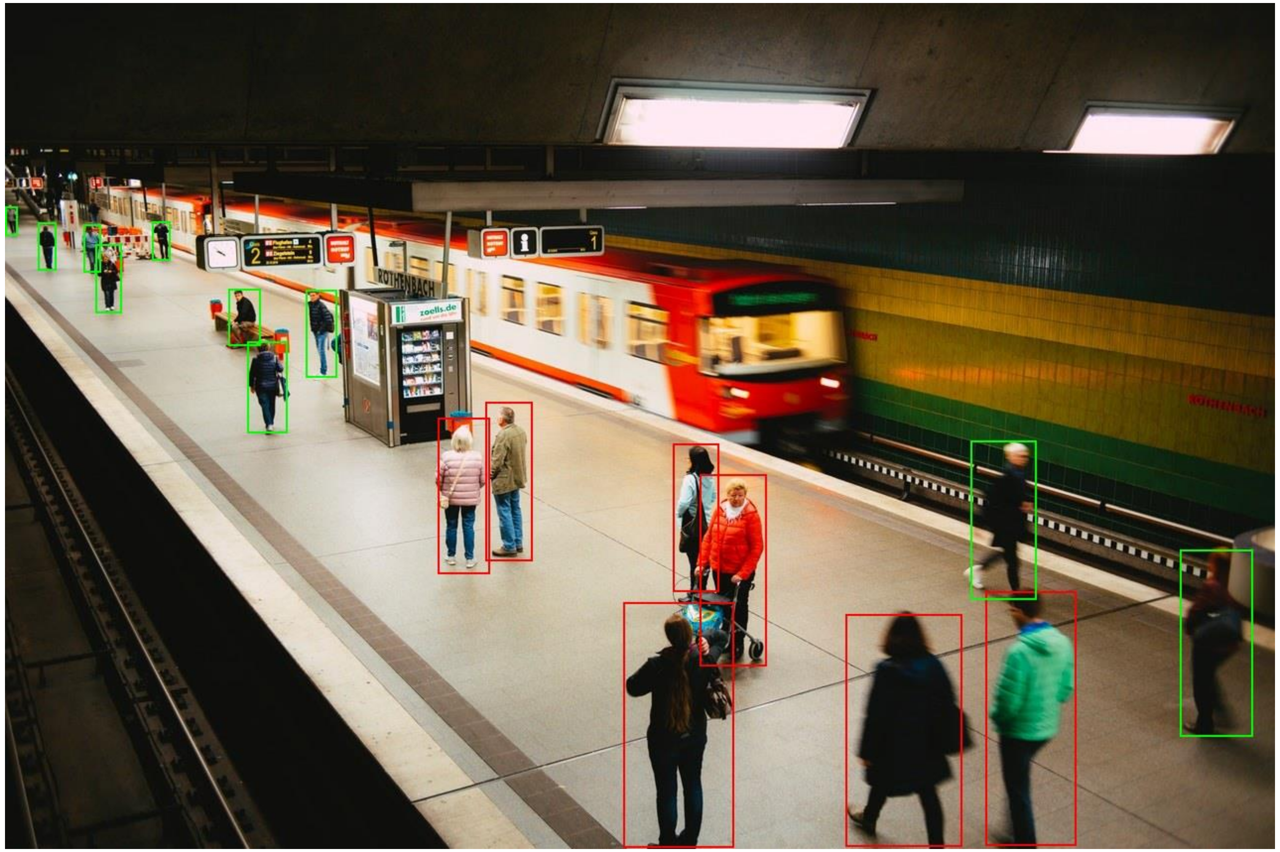
Object Detection

- Object detection combines classification with localization for multiple instances.
- Outputs 0-*N* bounding boxes and class scores for each box.



Applications of Object Detection

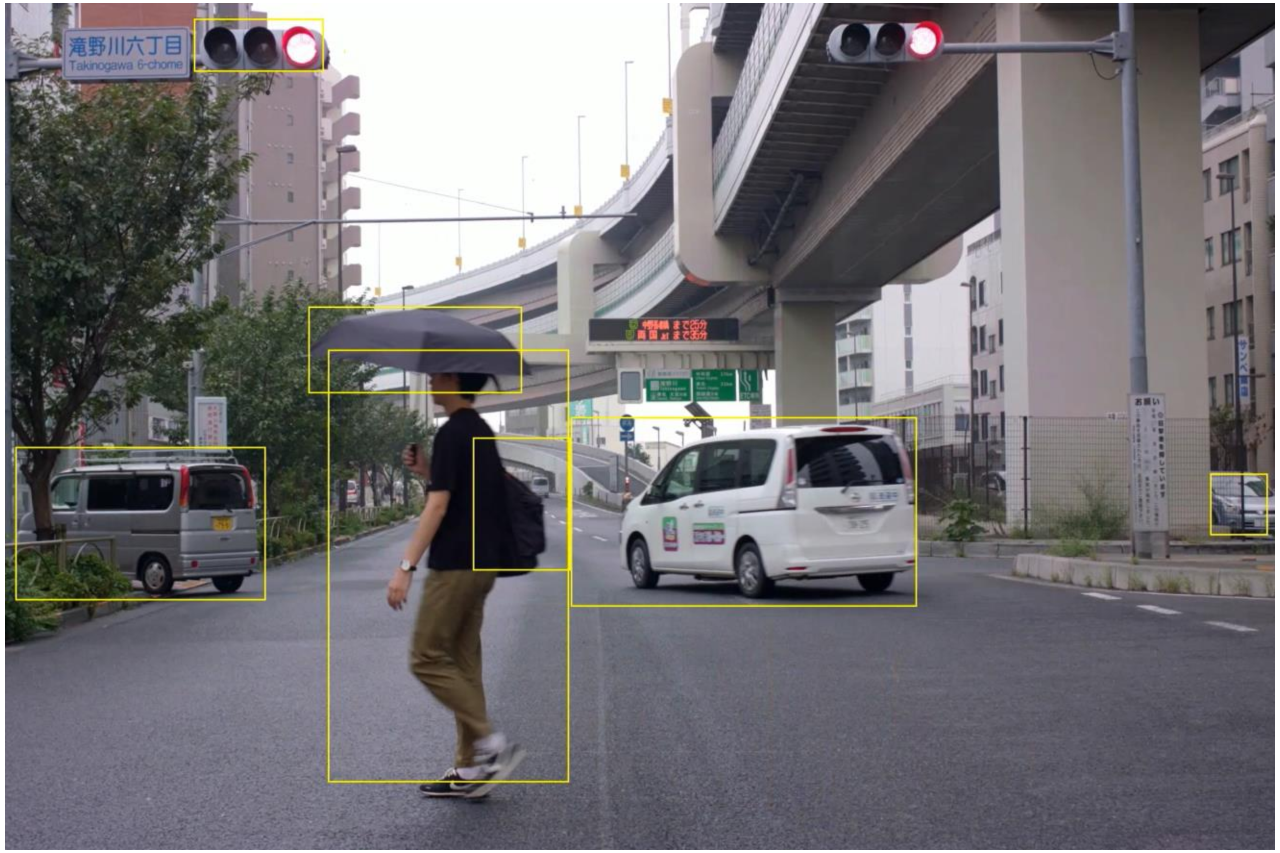
- Surveillance
  - Counting (cars, people, …)
  - + identification = recognition (faces, license plates, …)
  - + tracking = behavior analysis (hot spots, boundary crossing, package exchange, …)
- Autonomous vehicles



Pre-Trained Object Detection Models

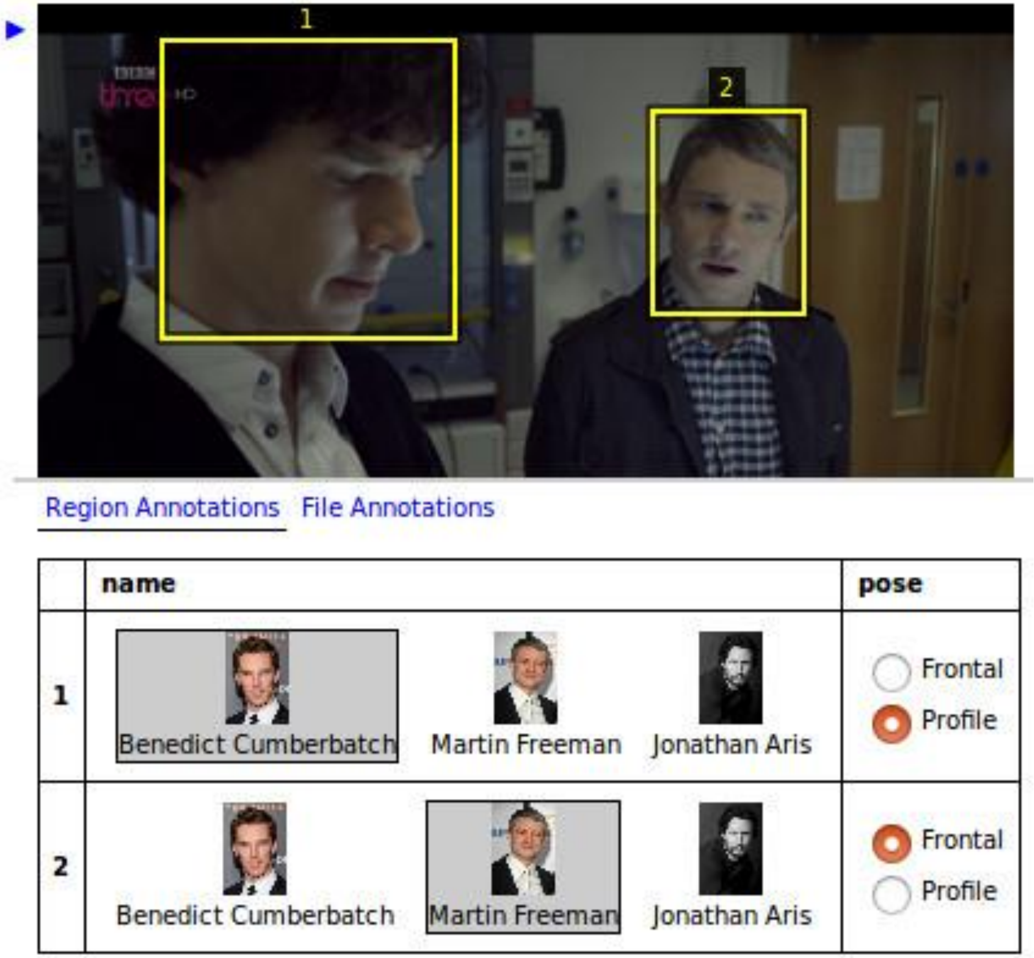
Public datasets with pre-trained models:

- COCO – 80 classes
- Pascal VOC – 20
- Pedestrians, faces, hands
- Cars, Bikes, Trains, …
- Street signs
- License plates
- Text



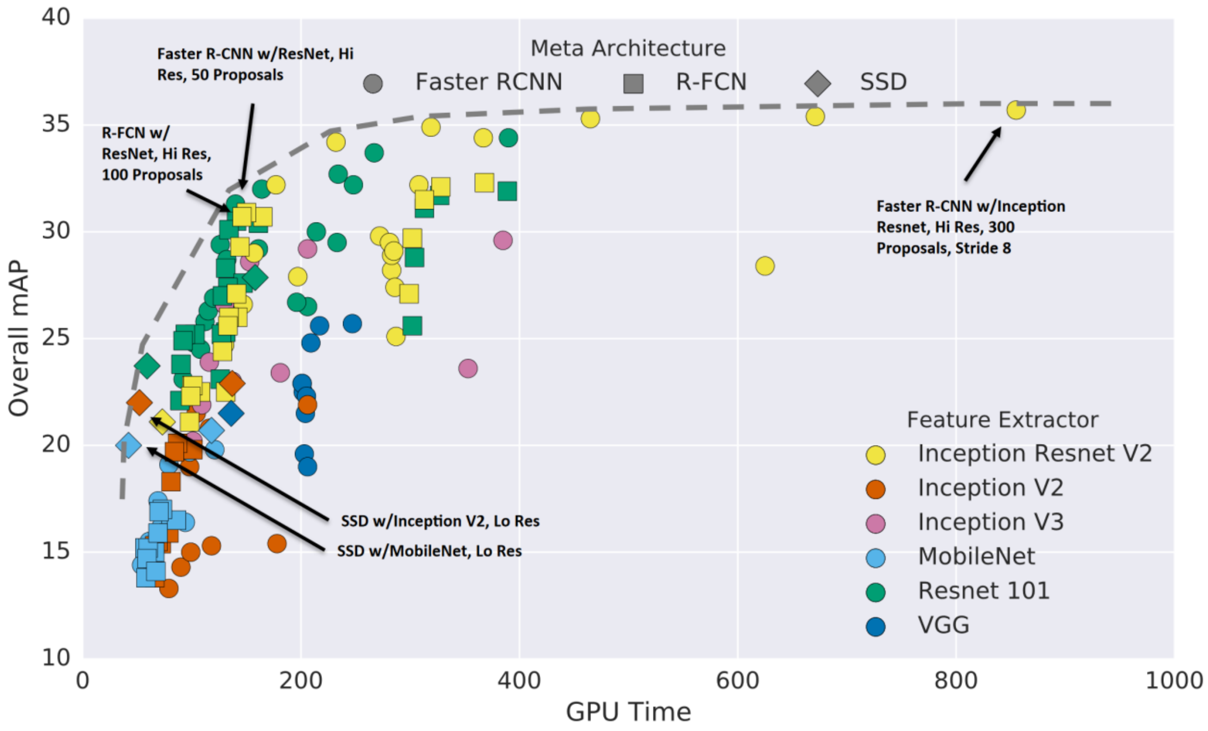
Training Data for Object Detection

- Labor intensive
- Labeling tool required
- Outline bounding box with mouse
- Label boxes with class
- Export and translate to suitable format



Object Detector Design Choices

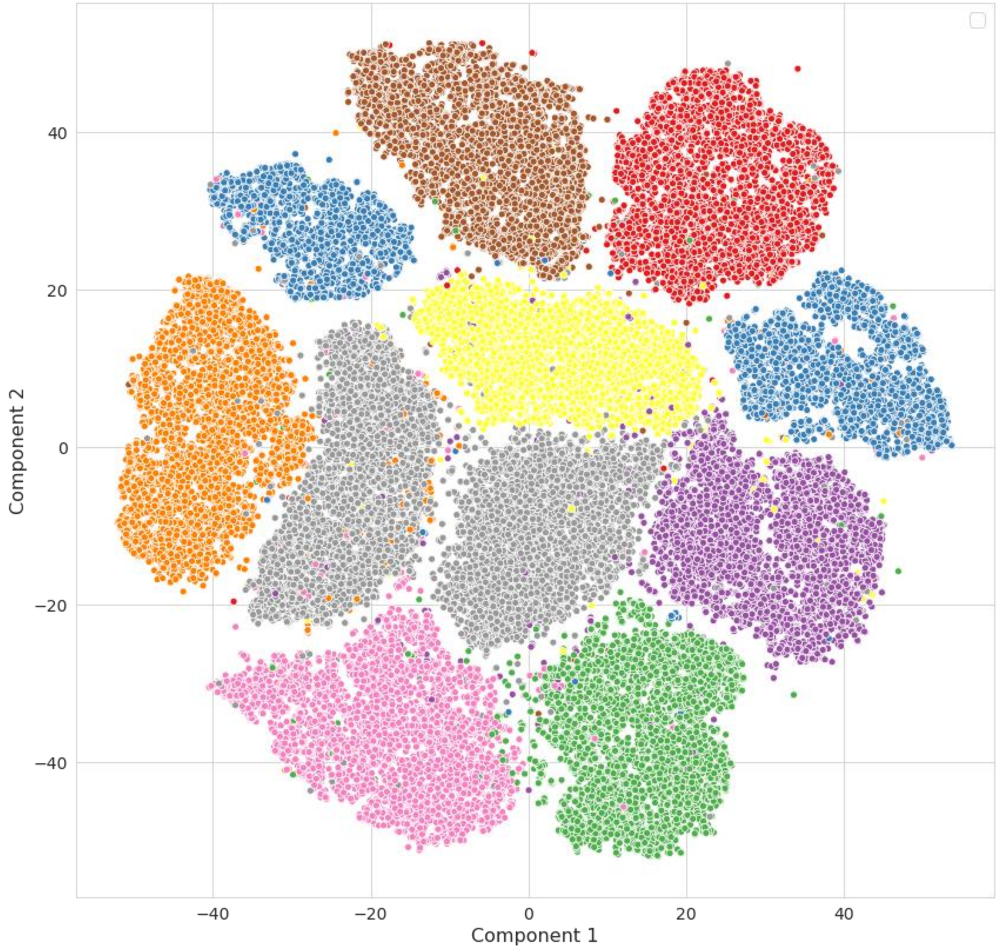
- Backbone (feature extractor)
- Meta-architecture (SSD, Yolo, Faster R-CNN)
- Number of proposals
- Resolution
- Training datasets



Huang, Jonathan, et al. "Speed/accuracy trade-offs for modern convolutional object detectors." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Embeddings

Embeddings



An embedding is a translation of a high-dimensional input to a low-dimensional feature vector that:

- captures semantics of the input, and
- preserves semantic similarity.

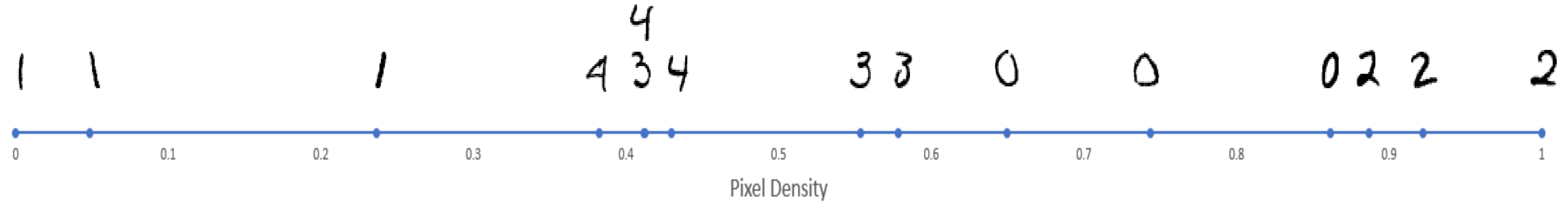
MNIST Digits Example

MNIST handwritten digit dataset

- 70,000 hand-written digits
- 28x28 grayscale images
- Each image is a point in $\mathbb{R}^{784}$

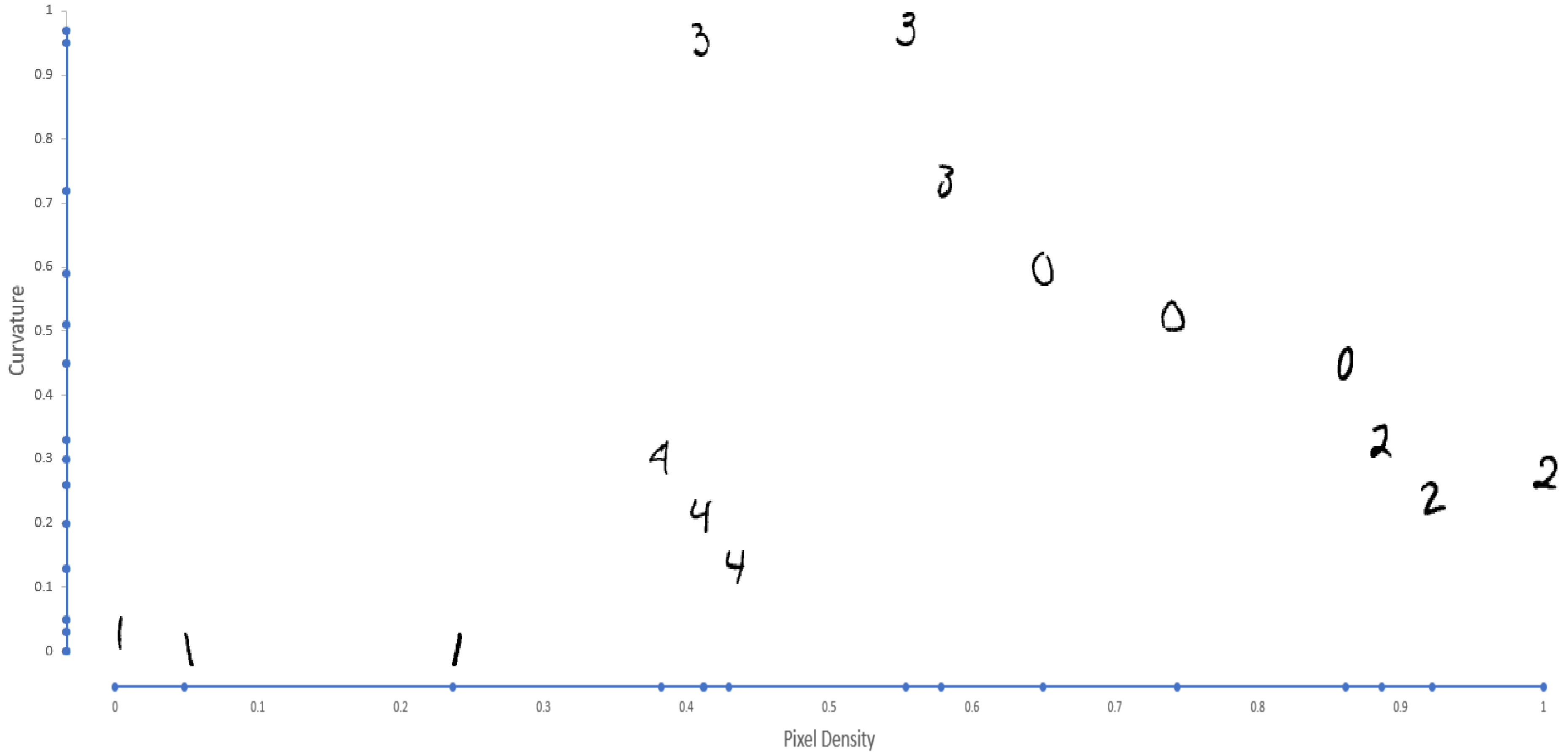
MNIST 1-D Embedding

Mean(image): $\mathbb{R}^{784} \rightarrow \mathbb{R}^{1}$



MNIST 2-D Embedding

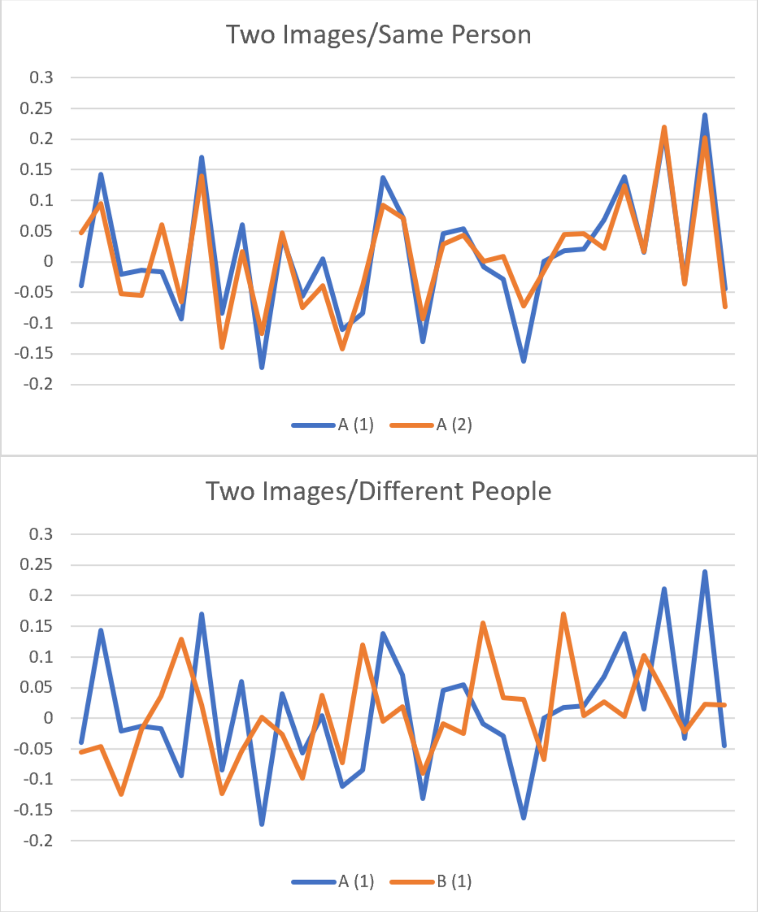
Mean(image),Curvature(image): $\mathbb{R}^{784} \rightarrow \mathbb{R}^2$



Facial Image Recognition

FaceNet Embedding:

- Deep Neural Network
- Face image $\rightarrow \mathbb{R}^{128}$
- Triplet loss function:
    - Anchor image
    - Positive image
    - Negative image



Conclusions

- Image Classification
  - 3 types: binary, multi-class, and multi-label
  - Answers questions about the whole image
- Objection Detection
  - Identifies multiple objects in an image and their locations
  - Major uses include counting, identification and object tracking
- Embeddings
  - Represents semantics of the input as a vector of real numbers
  - Used in facial image recognition technology

Resources

Papers

S. Bianco et al. 2018, **Benchmark Analysis of Representative Deep Neural Network Architectures**

https://arxiv.org/abs/1810.00736

J. Huang et al. 2017, **Speed/accuracy trade-offs for modern convolutional object detectors**

https://arxiv.org/abs/1611.10012

F. Schroff et al. 2015, **FaceNet: A Unified Embedding for Face Recognition and Clustering**

https://arxiv.org/abs/1503.03832

Pre-trained models

Tensorflow

https://tfhub.dev/

PyTorch

https://pytorch.org/hub/

Caffe

https://github.com/BVLC/caffe/wiki/Model-Zoo

Contact Information

thibault@streamlogic.io

https://streamlogic.io