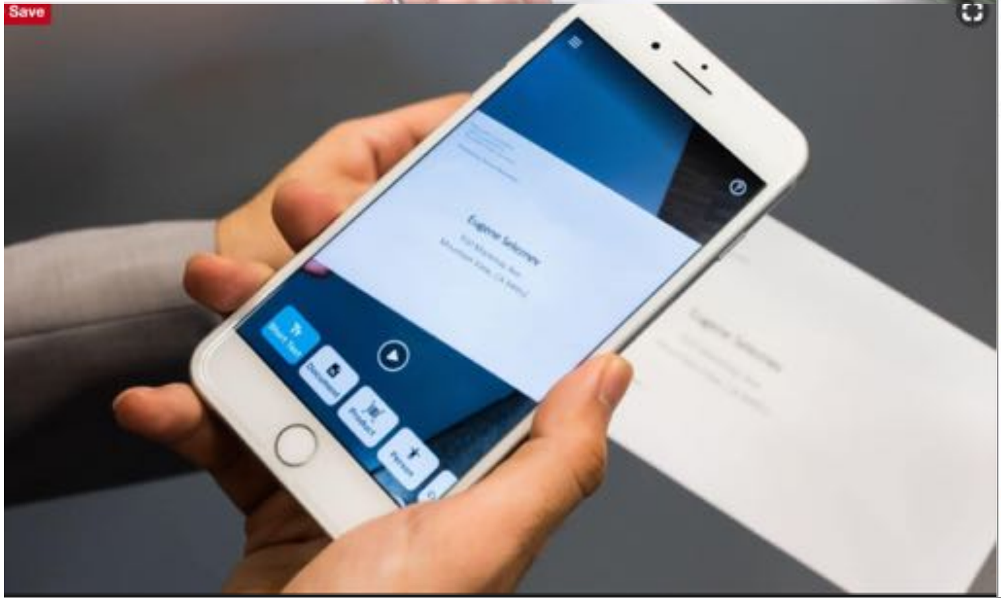
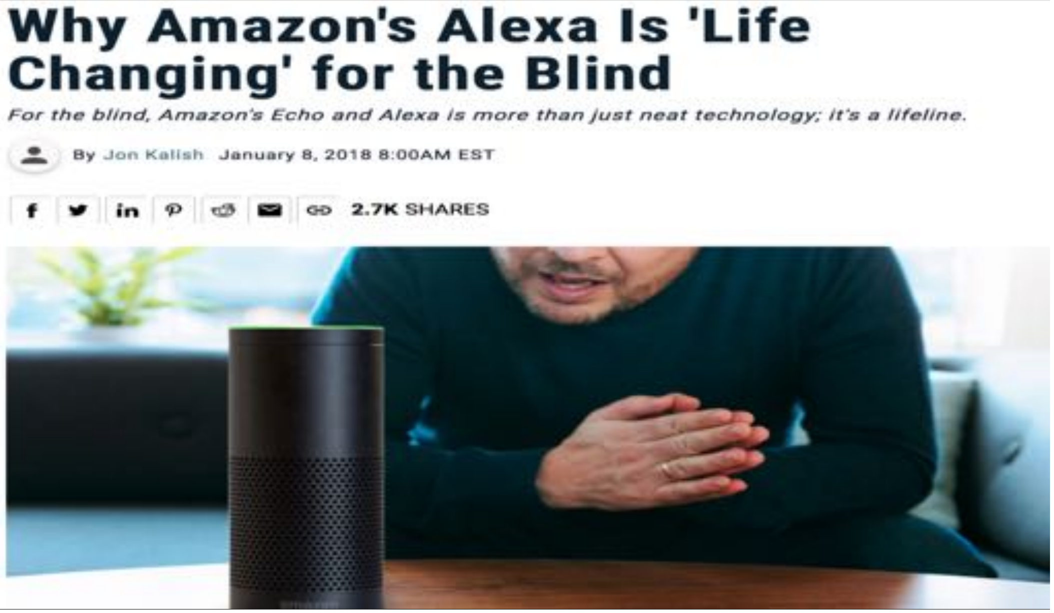
Apache MXNet (Incubating)

Steffen Rochel
steroche@amazon.com; @srochel



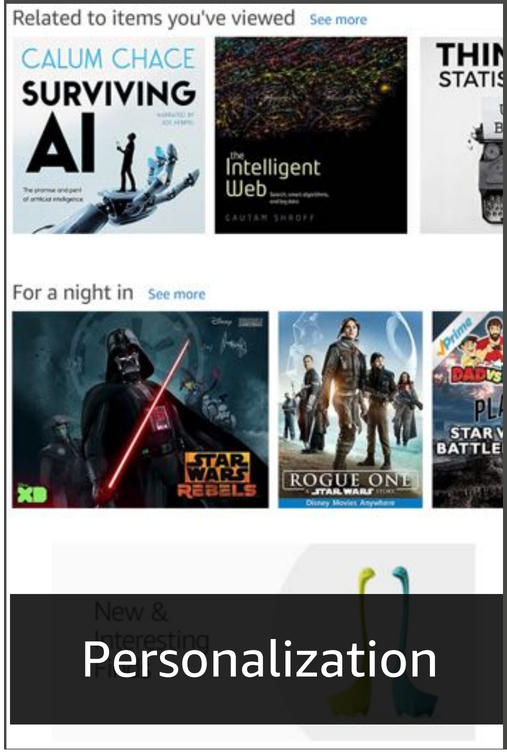


Why Amazon's Alexa Is 'Life Changing' for the Blind

For the blind, Amazon's Echo and Alexa is more than just neat technology; it's a lifeline.

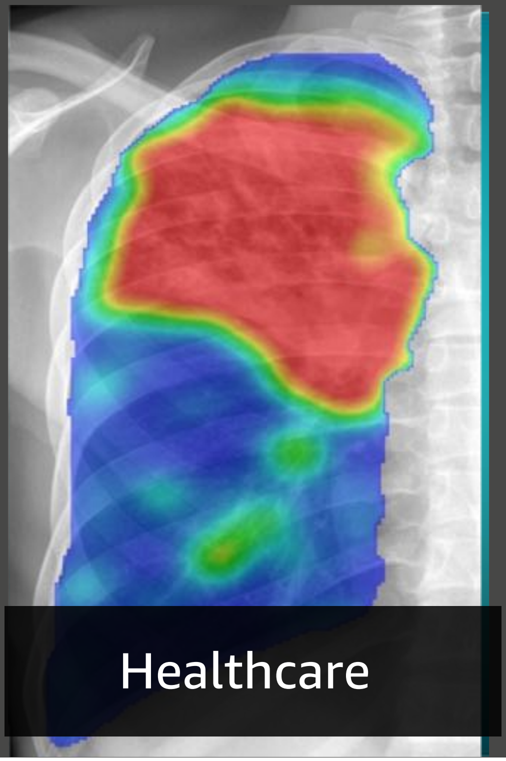
By Jon Kalish January 8, 2018 8:00AM EST

Why Deep Learning?

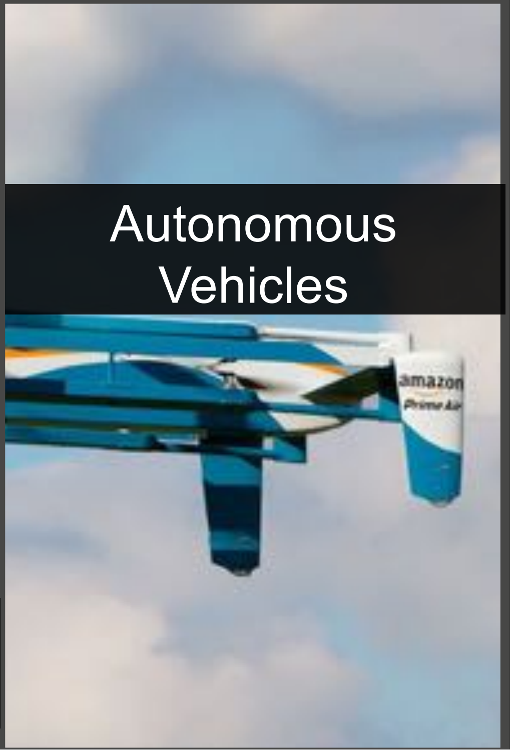


Personalization

Logistics

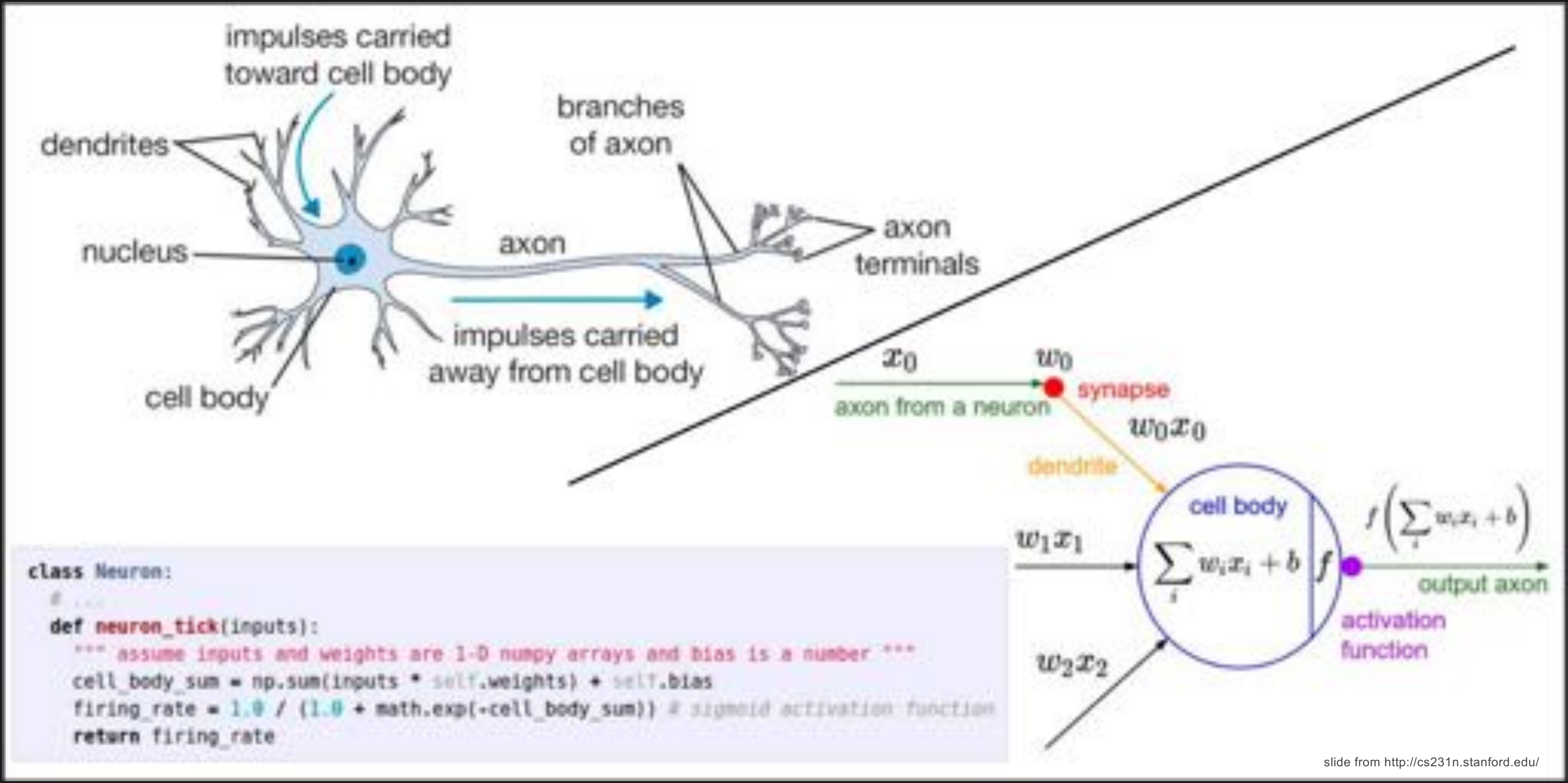


Healthcare



Autonomous Vehicles

Biological & Artificial Neuron

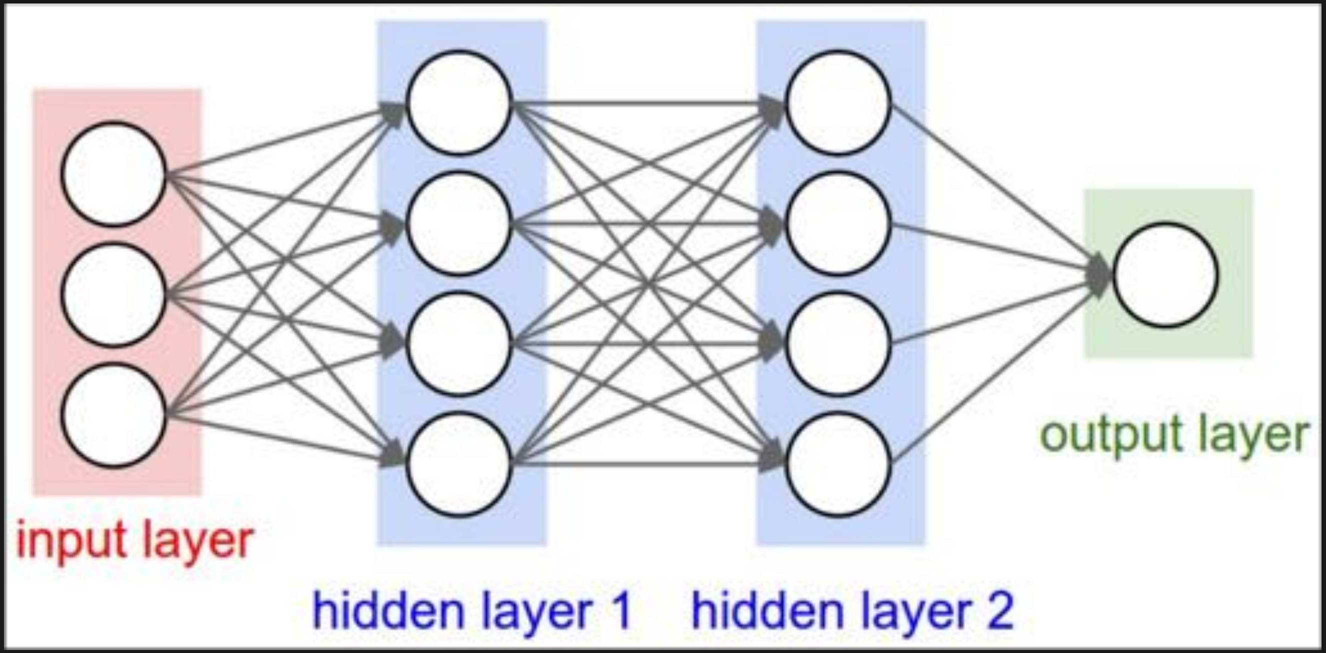


```
class Neuron:
  # ...
  def neuron_tick(inputs):
    """ assume inputs and weights are 1-D numpy arrays and bias is a number """
    cell_body_sum = np.sum(inputs * self.weights) + self.bias
    firing_rate = 1.0 / (1.0 + math.exp(-cell_body_sum)) # sigmoid activation function
    return firing_rate
```

slide from http://cs231n.stanford.edu/

Source: http://cs231n.github.io/neural-networks-1/

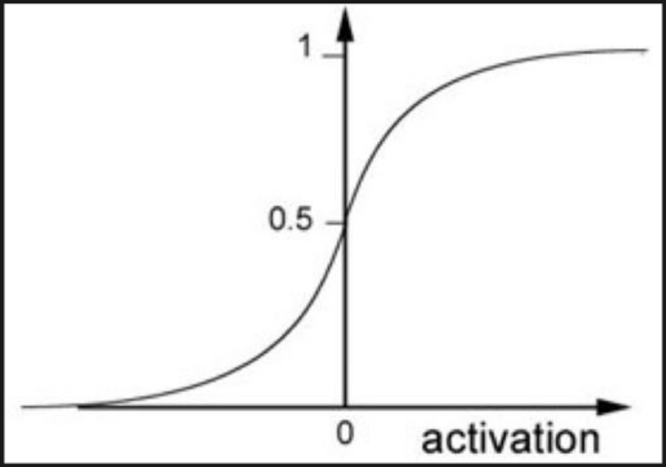
Fully Connected Layer

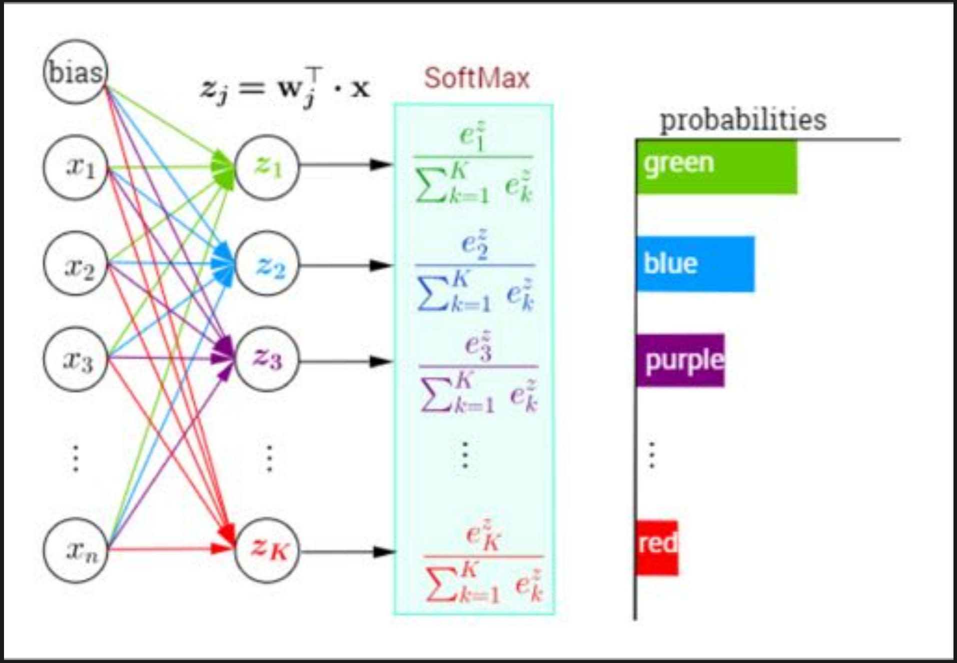


Each node (“neuron”) in a layer is connected to every node in the previous layer

Classification with the Softmax Function

Softmax Function

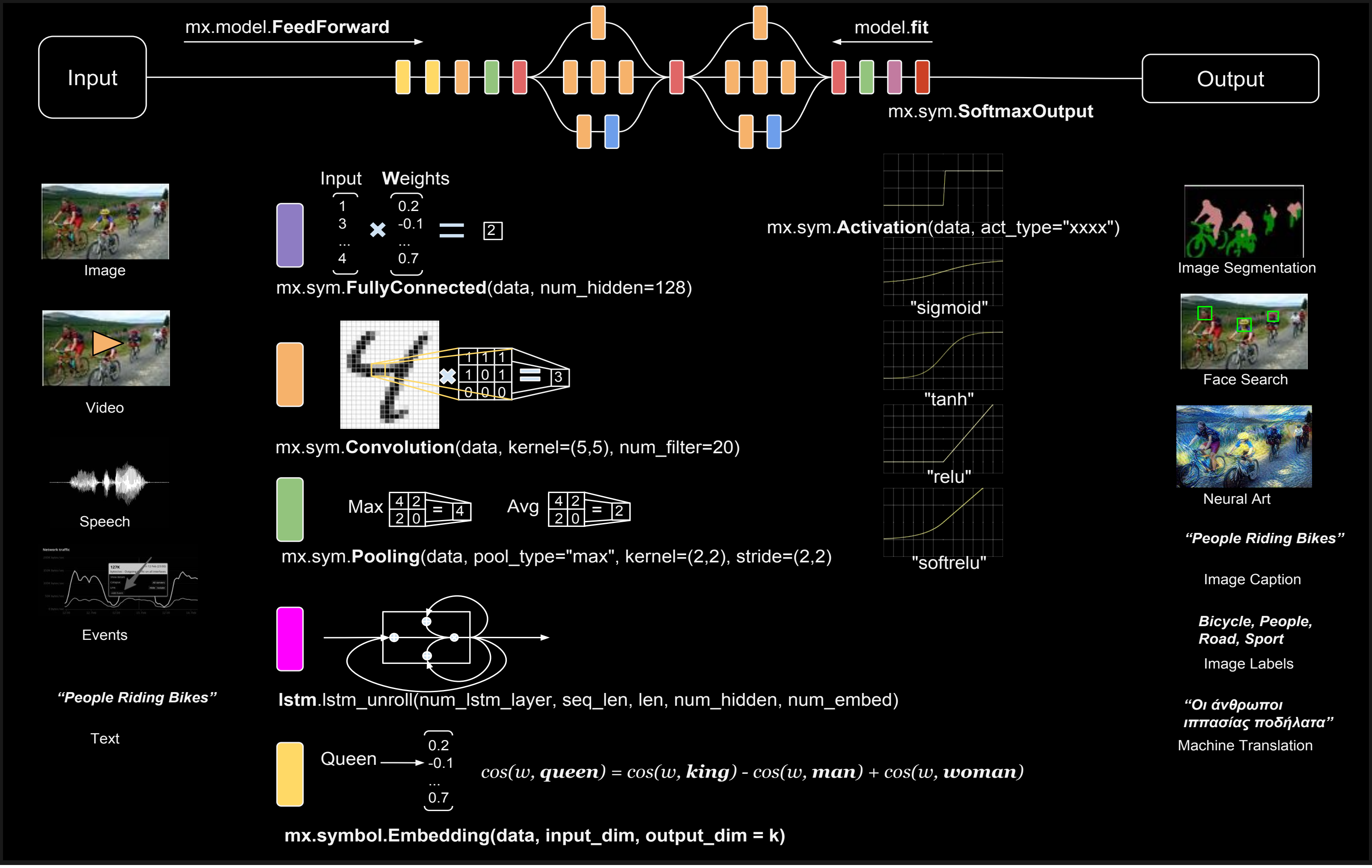



bias
x_1
x_2
x_3
x_n
$z_j = \mathbf{w}_j^\top \cdot \mathbf{x}$
z_1
z_2
z_3
z_K
SoftMax
$\frac{e^{z}_1}{\sum_{k=1}^K e^{z}_k}$
$\frac{e^{z}_2}{\sum_{k=1}^K e^{z}_k}$
$\frac{e^{z}_3}{\sum_{k=1}^K e^{z}_k}$
$\frac{e^{z}_K}{\sum_{k=1}^K e^{z}_k}$
probabilities
green
blue
purple
red

Softmax converts the output layer into probabilities – necessary for classification

Source: https://stats.stackexchange.com/questions/273465/neural-network-softmax-activation

Deep Learning Models



Input

mx.model.**FeedForward**

model.**fit**

Output

mx.sym.**SoftmaxOutput**

Image

Video

Speech

Events

"People Riding Bikes"

Text

mx.sym.**FullyConnected**(data, num_hidden=128)

mx.sym.**Convolution**(data, kernel=(5,5), num_filter=20)

mx.sym.**Pooling**(data, pool_type="max", kernel=(2,2), stride=(2,2))

lstm.lstm_unroll(num_lstm_layer, seq_len, len, num_hidden, num_embed)

$cos(w, \textbf{queen}) = cos(w, \textbf{king}) - cos(w, \textbf{man}) + cos(w, \textbf{woman})$

mx.symbol.Embedding(data, input_dim, output_dim = k)

mx.sym.**Activation**(data, act_type="xxxx")

"sigmoid"

"tanh"

"relu"

"softrelu"

Image Segmentation

Face Search

Neural Art

"People Riding Bikes"

Image Caption

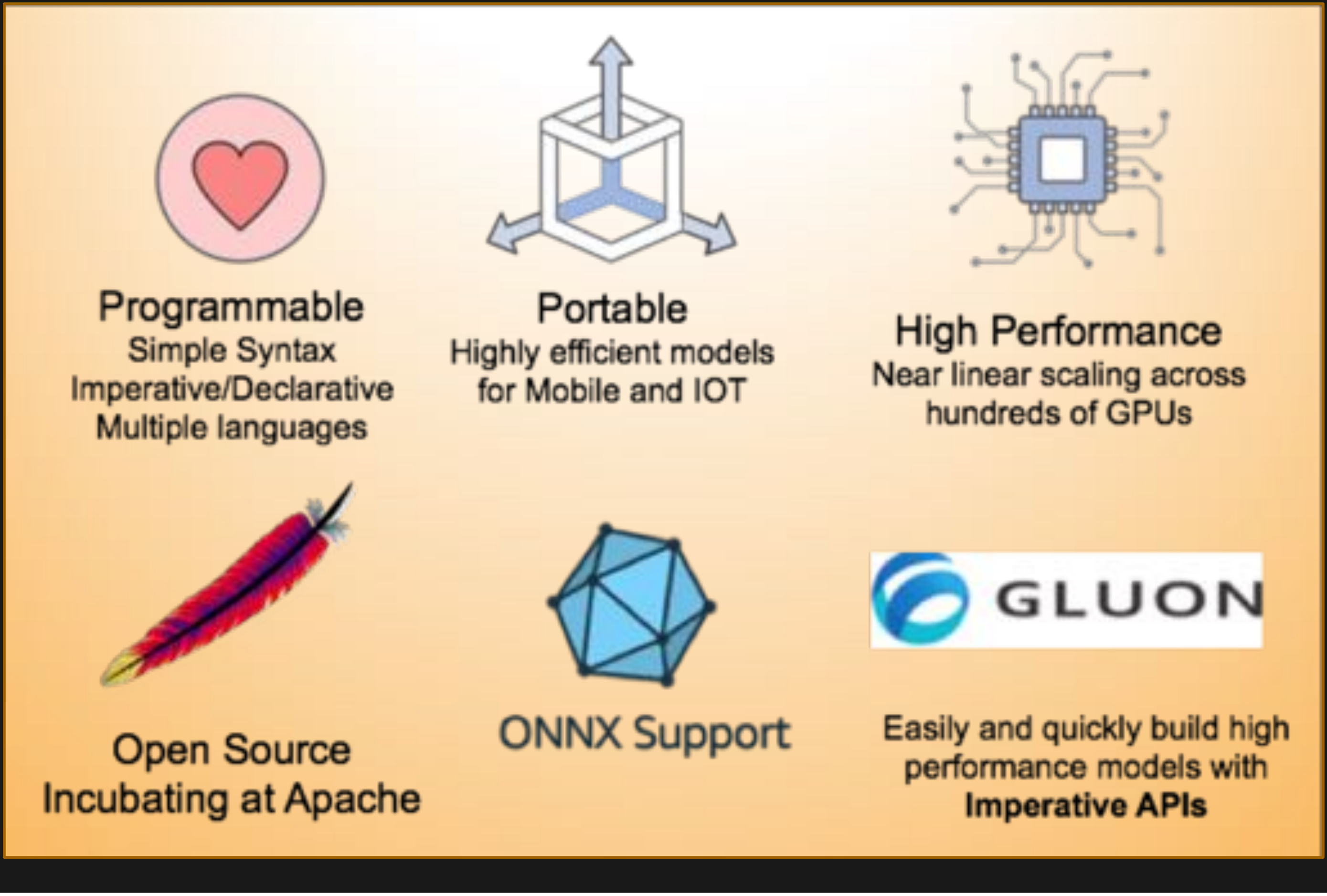
Bicycle, People, Road, Sport

Image Labels

"Οι άνθρωποι ιππασίας ποδήλατα"

Machine Translation

MXNet – a Deep Learning Framework



Programmable
Simple Syntax
Imperative/Declarative
Multiple languages

Portable
Highly efficient models
for Mobile and IOT

High Performance
Near linear scaling across
hundreds of GPUs

Open Source
Incubating at Apache

ONNX Support

GLUON

Easily and quickly build high
performance models with
Imperative APIs

Easily build, train, and deploy models with MXNet

Start with high quality, pre-trained models

- Gluon CV and Gluon NLP

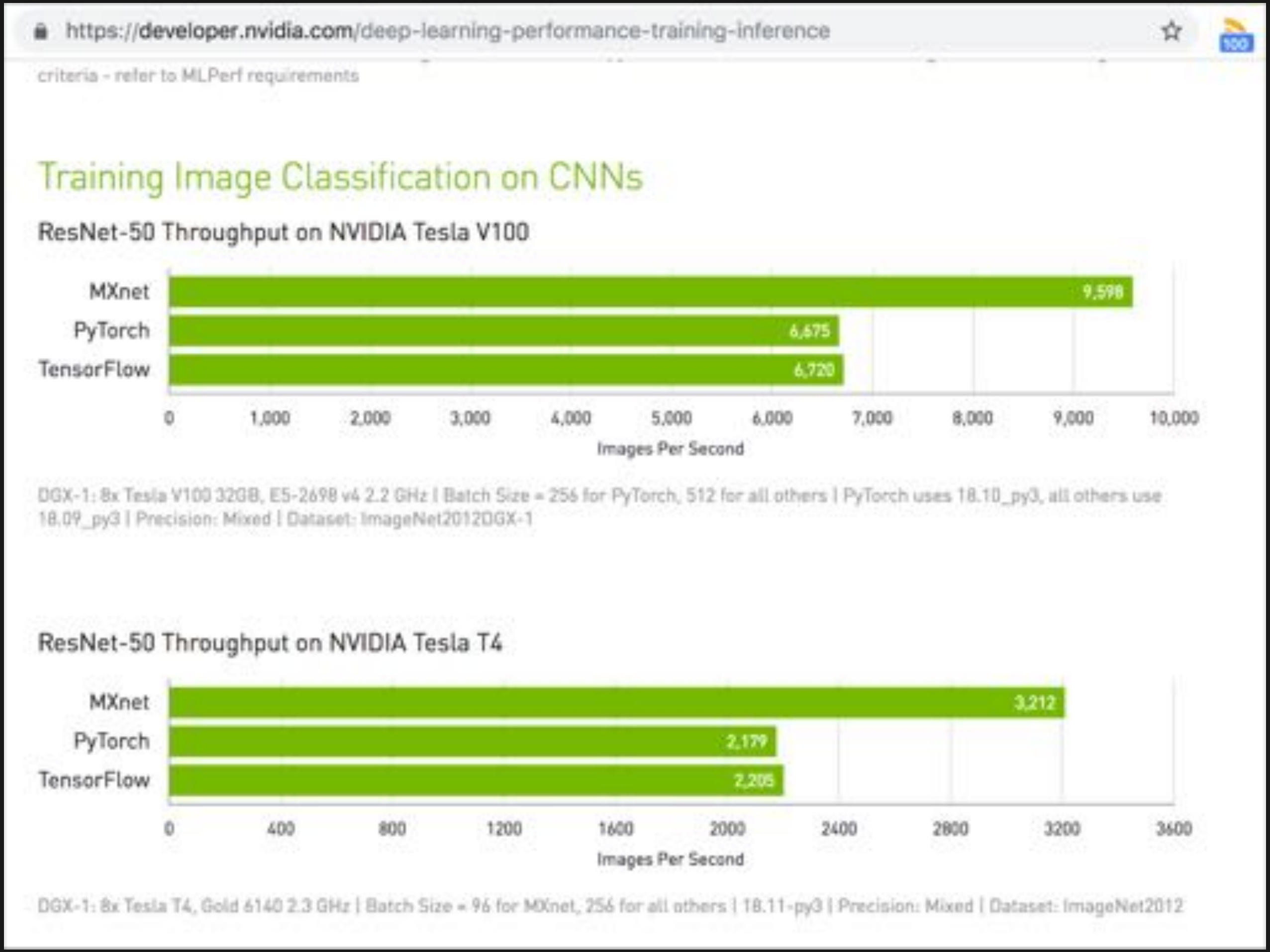
Refine with fast, scalable training

- Keras-MXNet up to 2x faster than Keras-TensorFlow
- Near-linear scalability up to 256 GPUs
- Dynamic training

Deploy using familiar tools

- Java/Scala APIs
- MXNet Model Server

Performance



https://developer.nvidia.com/deep-learning-performance-training-inference

criteria - refer to MLPerf requirements

Training Image Classification on CNNs

ResNet-50 Throughput on NVIDIA Tesla V100

DGX-1: 8x Tesla V100 32GB, E5-2698 v4 2.2 GHz | Batch Size = 256 for PyTorch, 512 for all others | PyTorch uses 18.10_py3, all others use 18.09_py3 | Precision: Mixed | Dataset: ImageNet2012DGX-1

ResNet-50 Throughput on NVIDIA Tesla T4

DGX-1: 8x Tesla T4, Gold 6140 2.3 GHz | Batch Size = 96 for MXnet, 256 for all others | 18.11-py3 | Precision: Mixed | Dataset: ImageNet2012

Why Gluon

Simple, Easy-to-Understand Code

Flexible, Imperative Structure

Dynamic Graphs

High Performance

Gluon Code – Network Definition

```
net = gluon.nn.HybridSequential()

with net.name_scope():

    net.add(gluon.nn.Dense(units=64, activation='relu'))

    net.add(gluon.nn.Dense(units=10))

softmax_cross_entropy = gluon.loss.SoftmaxCrossEntropyLoss()

net.initialize(mx.init.Xavier(magnitude=2.24), ctx=ctx, force_reinit=True)

trainer = gluon.Trainer(net.collect_params(), 'sgd', {'learning_rate': 0.02})
```

Gluon Code – Training

```
smoothing_constant = .01
for e in range(10):
    cumulative_loss = 0
    for i, (data, label) in enumerate(train_data):
        data = data.as_in_context(model_ctx).reshape((-1, 784))
        label = label.as_in_context(model_ctx)
        with autograd.record():
            output = net(data)
            loss = softmax_cross_entropy(output, label)
        loss.backward()
        trainer.step(data.shape[0])
```

MXNet EcoSystem

- Gluon Model Zoo
  https://mxnet.incubator.apache.org/api/python/gluon/model_zoo.html
- Sockeye: A Toolkit for Neural Machine Translation
  https://arxiv.org/abs/1712.05690
- GluonCV: A Deep Learning Toolkit for Computer Vision
  https://gluon-cv.mxnet.io/
- GluonNLP: a Deep Learning Toolkit for Natural Language Processing
  http://gluon-nlp.mxnet.io/
- DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks
  https://arxiv.org/abs/1704.04110v2
- MXNet Model Server
  https://github.com/awslabs/mxnet-model-server

MXNET.IO

https://www.meetup.com/Deep-Learning-with-Apache-MXNet-Berlin/

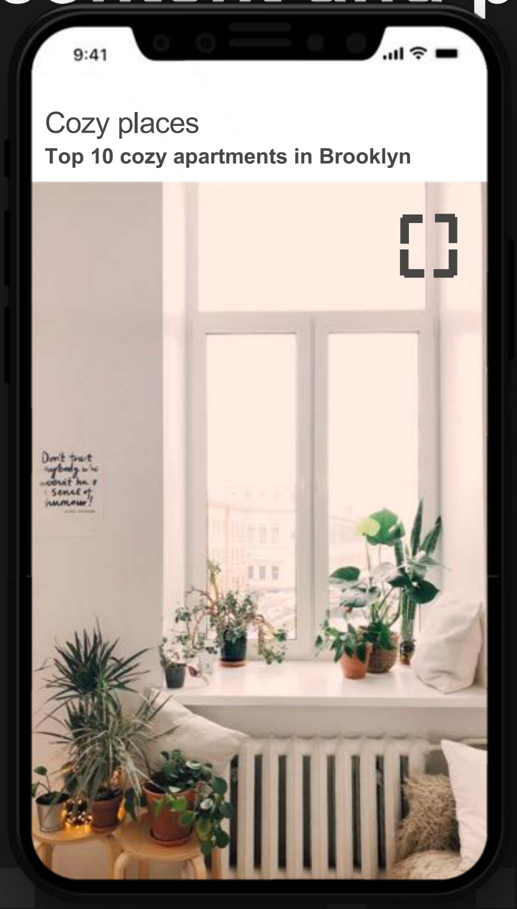
Steffen Rochel
@srochel

Visual Search

AWS re:Invent

aws

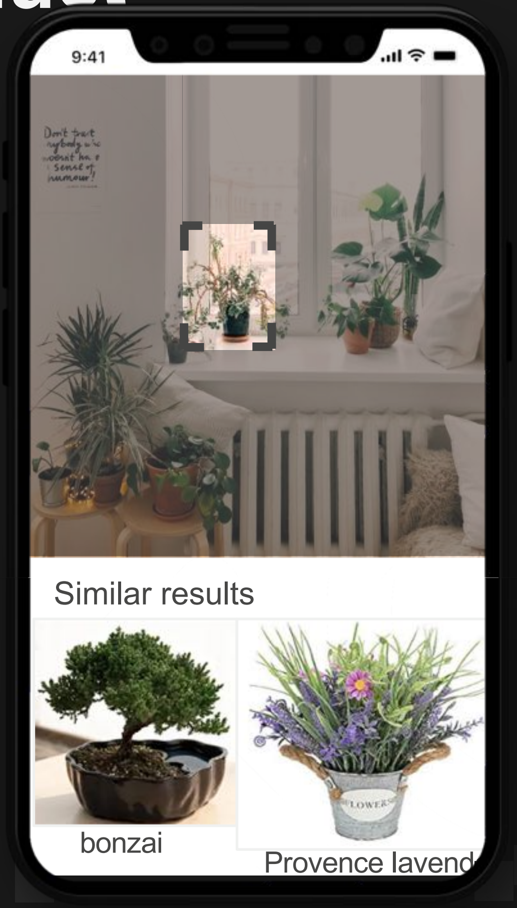
Drive content and product



9:41

Cozy places

Top 10 cozy apartments in Brooklyn



9:41

Similar results

bonzai

Provence lavend

Search Engines

Bing search

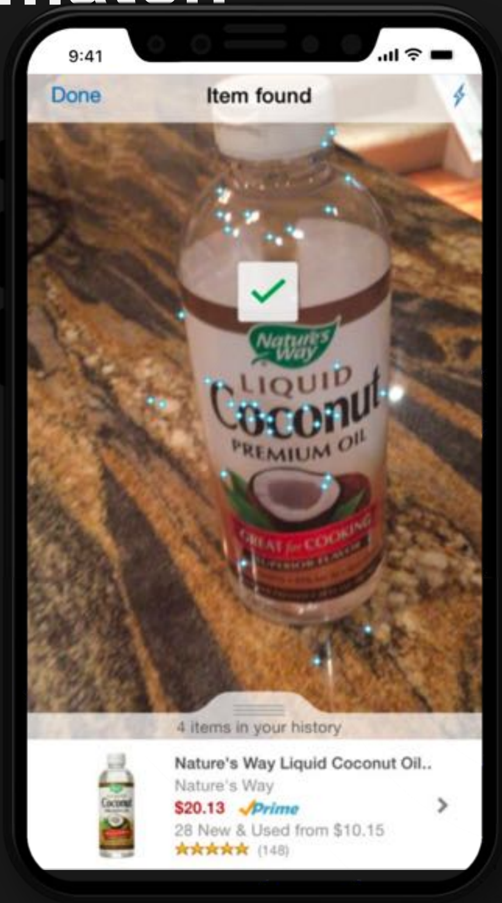
Google search

Social

Pinterest

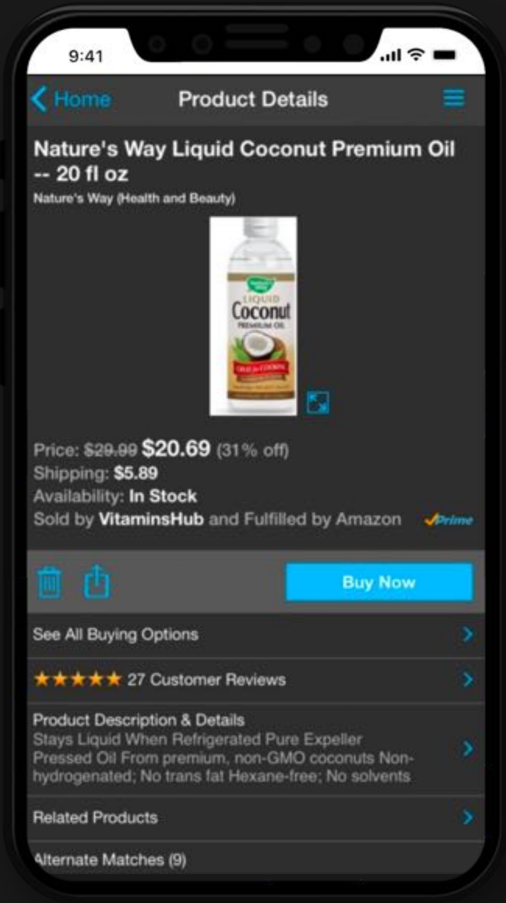
(not actual screenshot)

Exact match



4 items in your history

Nature's Way Liquid Coconut Oil..
Nature's Way
$20.13 ✓Prime
28 New & Used from $10.15
★★★★★ (148)



Nature's Way Liquid Coconut Premium Oil -- 20 fl oz

Nature's Way (Health and Beauty)

Price: ~~$29.99~~ $20.69 (31% off)
Shipping: $5.89
Availability: In Stock
Sold by VitaminsHub and Fulfilled by Amazon ✓Prime

See All Buying Options ›

★★★★★ 27 Customer Reviews ›

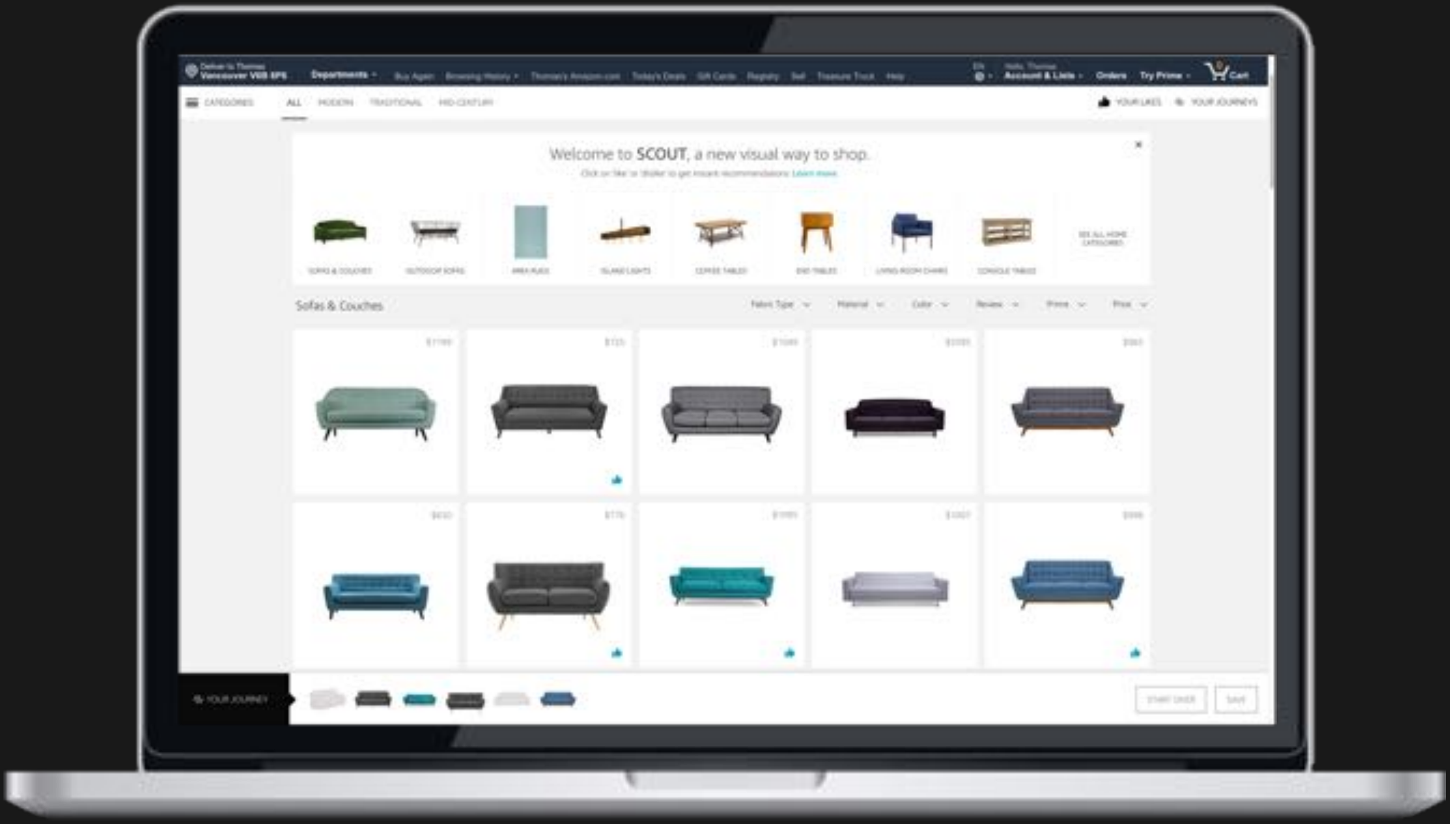
Product Description & Details
Stays Liquid When Refrigerated Pure Expeller Pressed Oil From premium, non-GMO coconuts Non-hydrogenated; No trans fat Hexane-free; No solvents ›

Related Products ›

Alternate Matches (9)

Amazon Shopping

Visual recommendations



Welcome to SCOUT, a new visual way to shop.

Click on 'like' or 'dislike' to get instant recommendations. Learn more

Amazon **Scout**

> “The future of search will be about pictures rather than keywords.”

Ben Silbermann
Pinterest CEO

Growing market

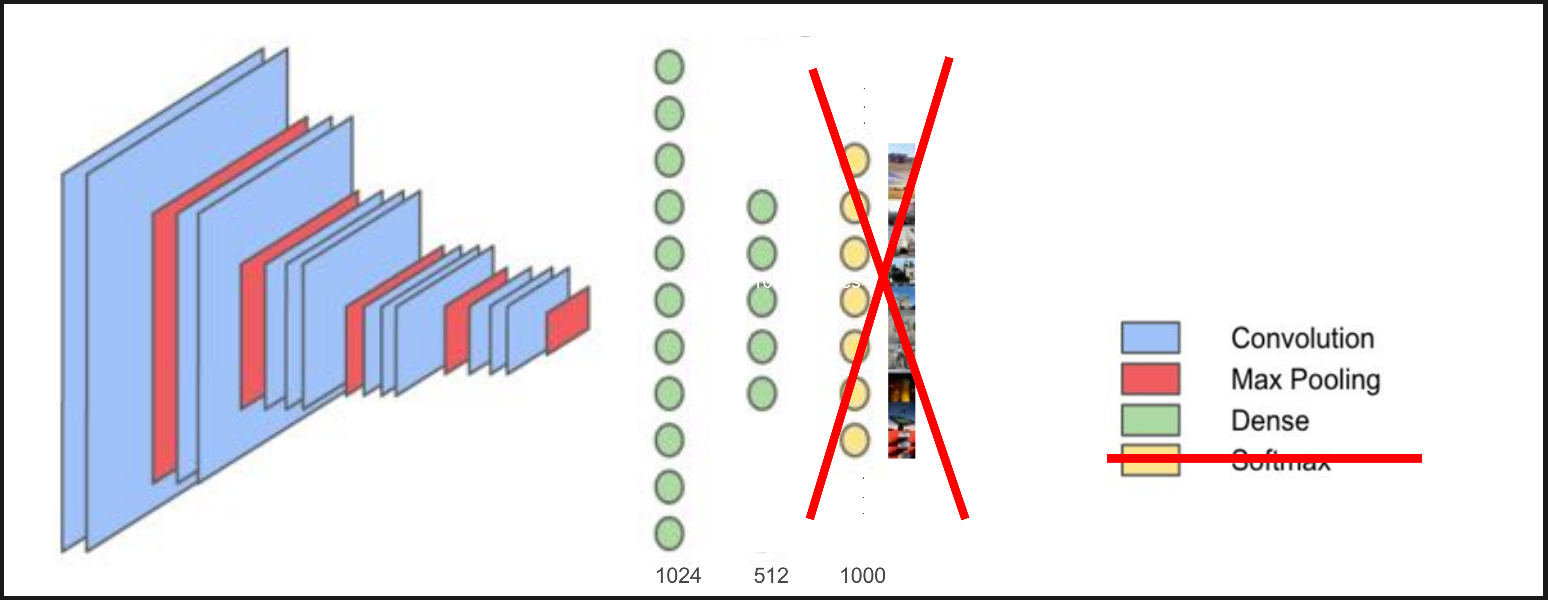
Angel.co: 76 startups

Syte.ai, Slyce.it and others

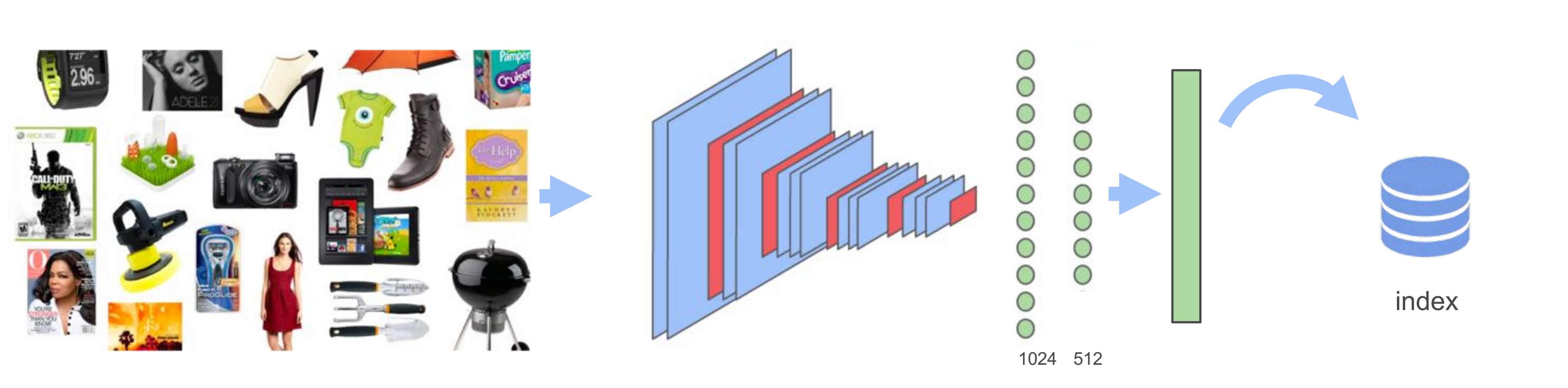
How does it work?

Turn the network into a featurizer

pre-trained CNN (resnet18) from the MXNet Gluon model zoo



Index images to lower-dimensional representation



Images: 224x224x3 dimensions

pre-trained network

512 dimensions “finger-print”

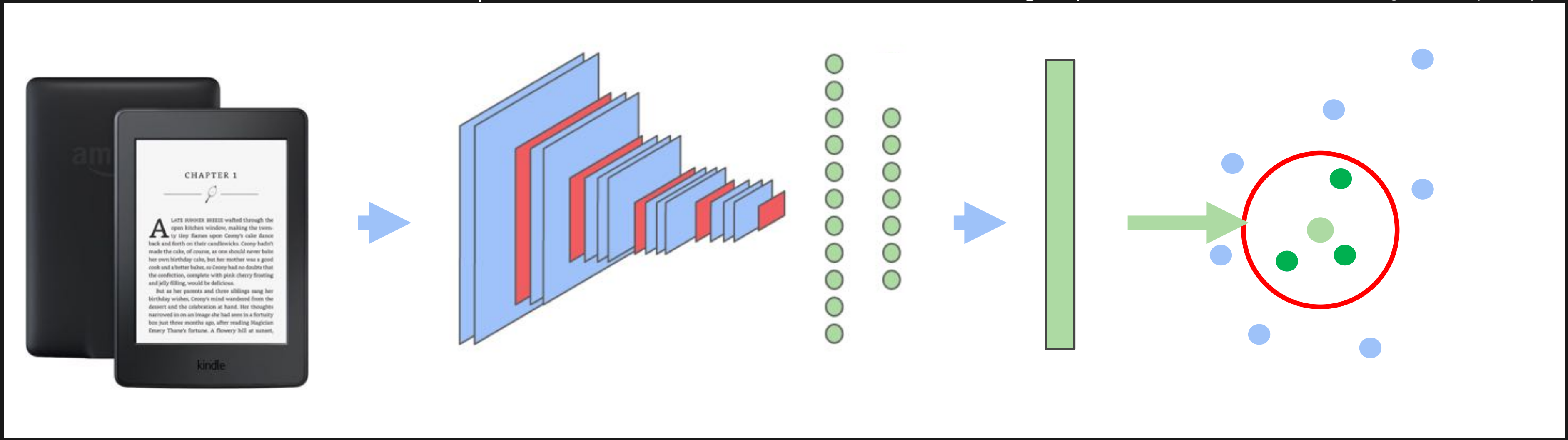
New image query

224x224x3 dimensions

pre-trained CNN

512 dimensions “finger-print”

k-Nearest Neighbor (k=3)

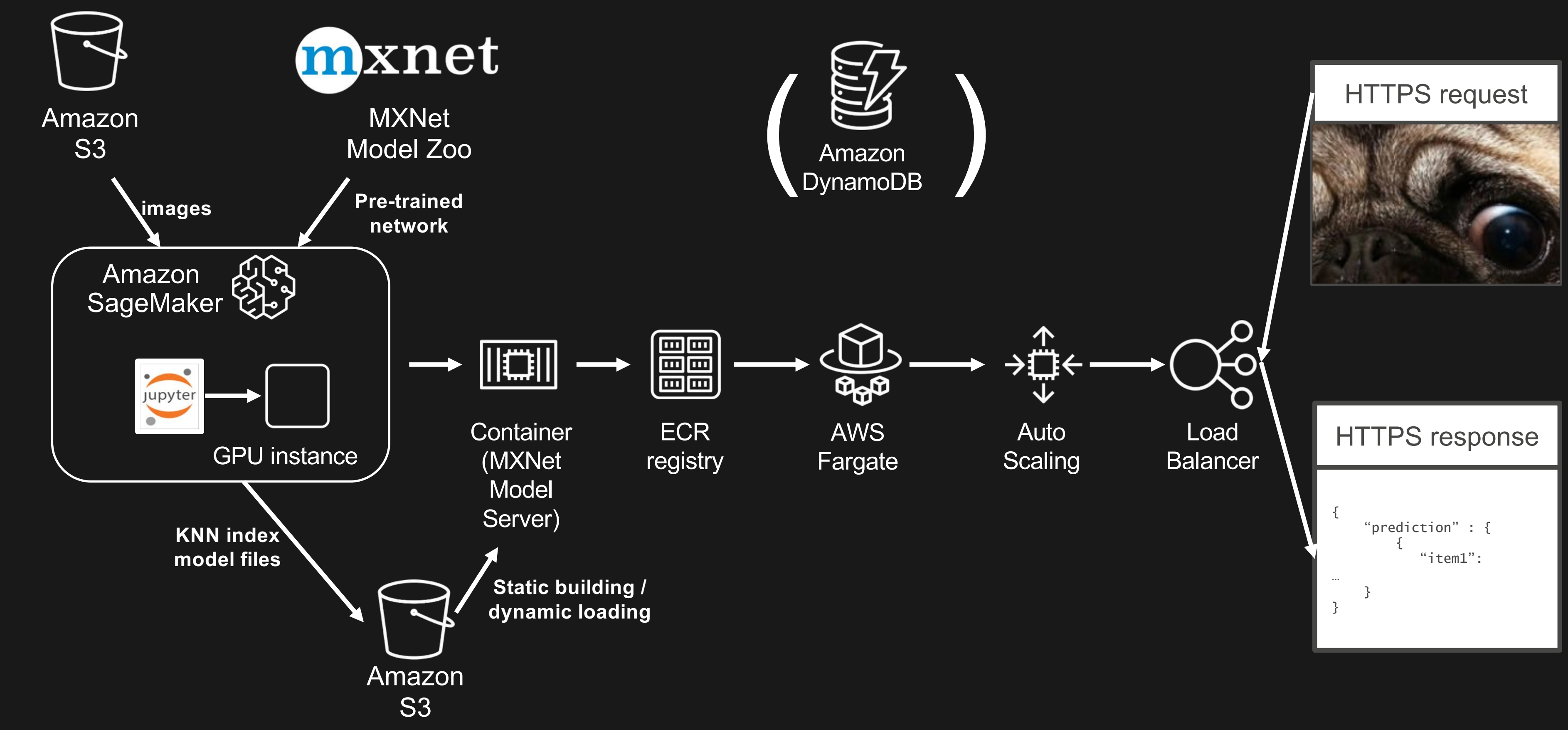


Implementation

AWS re:Invent

aws

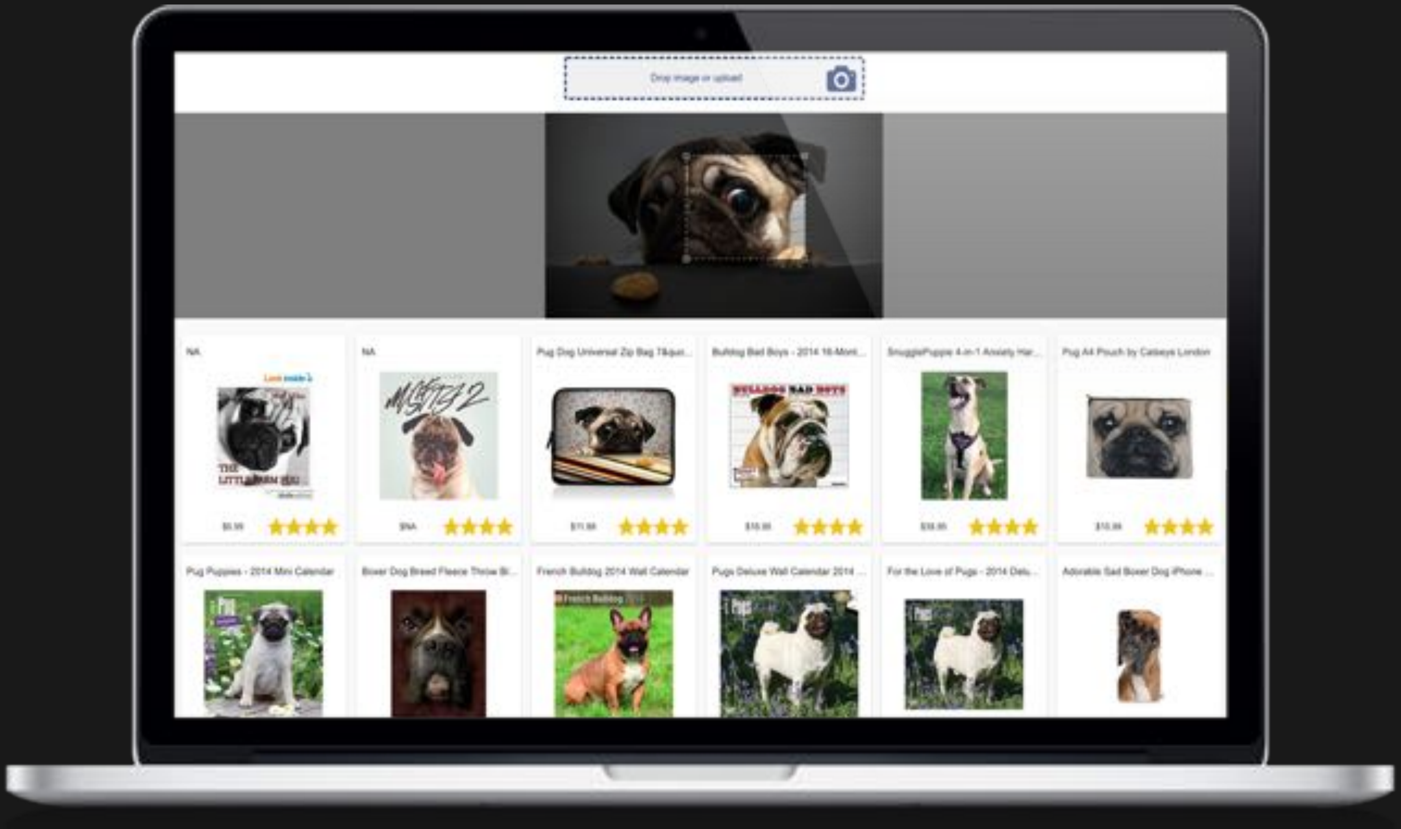
Workflow and operationalization



```
{
    “prediction” : {
        {
            “item1”:
…
    }
}
```

Demo

https://thomasdelteil.github.io/VisualSearch_MXNet/



Drop image or upload

NA

$5.99

NA

$NA

Pug Dog Universal Zip Bag 7&quot...

$11.98

Bulldog Bad Boys - 2014 16-Mont...

$15.99

SnugglePuppie 4-in-1 Anxiety Har...

$38.95

Pug A4 Pouch by Catseye London

$10.99

Pug Puppies - 2014 Mini Calendar

Boxer Dog Breed Fleece Throw Bl...

French Bulldog 2014 Wall Calendar

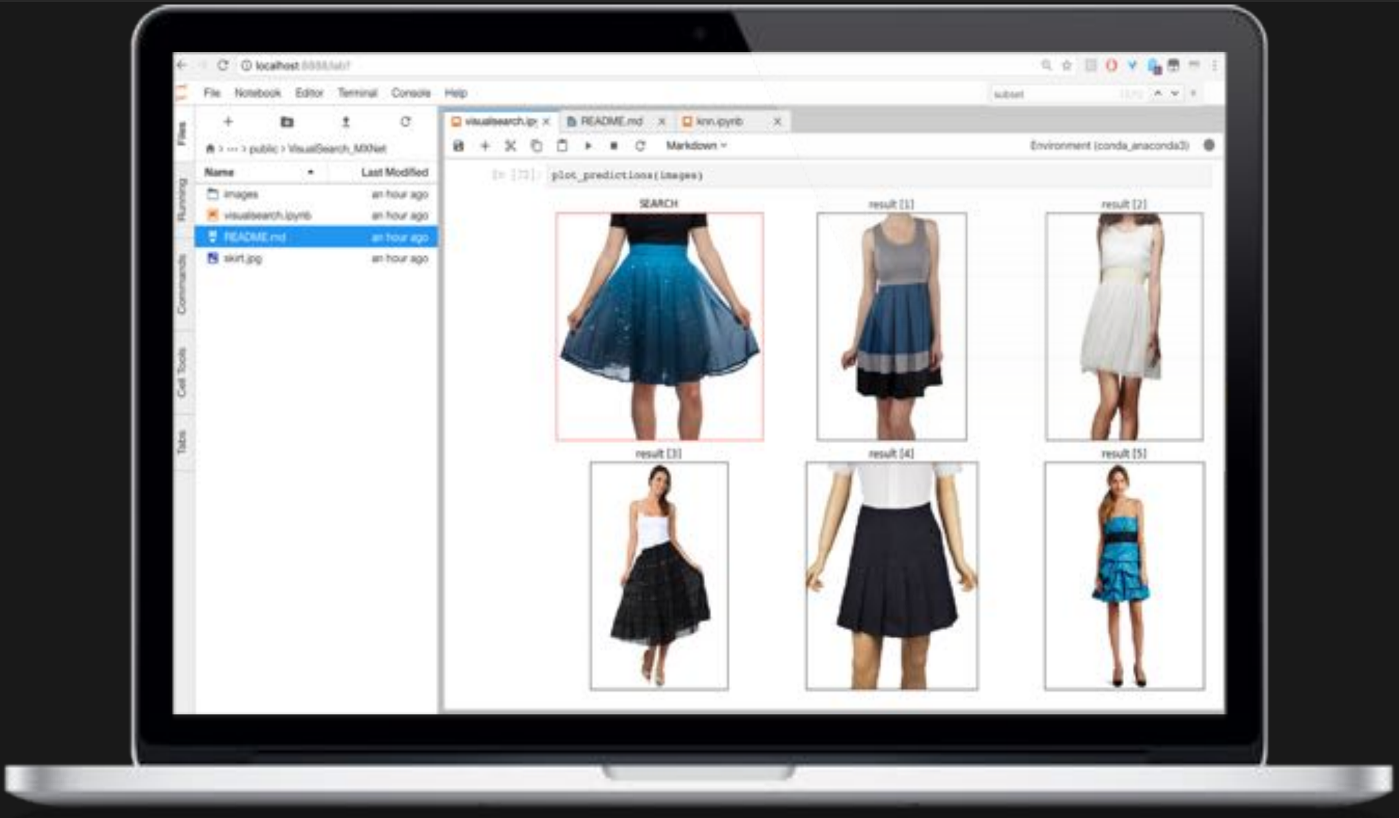
Pugs Deluxe Wall Calendar 2014 ...

For the Love of Pugs - 2014 Delu...

Adorable Sad Boxer Dog iPhone ...

Code

https://github.com/ThomasDelteil/VisualSearch_MXNet



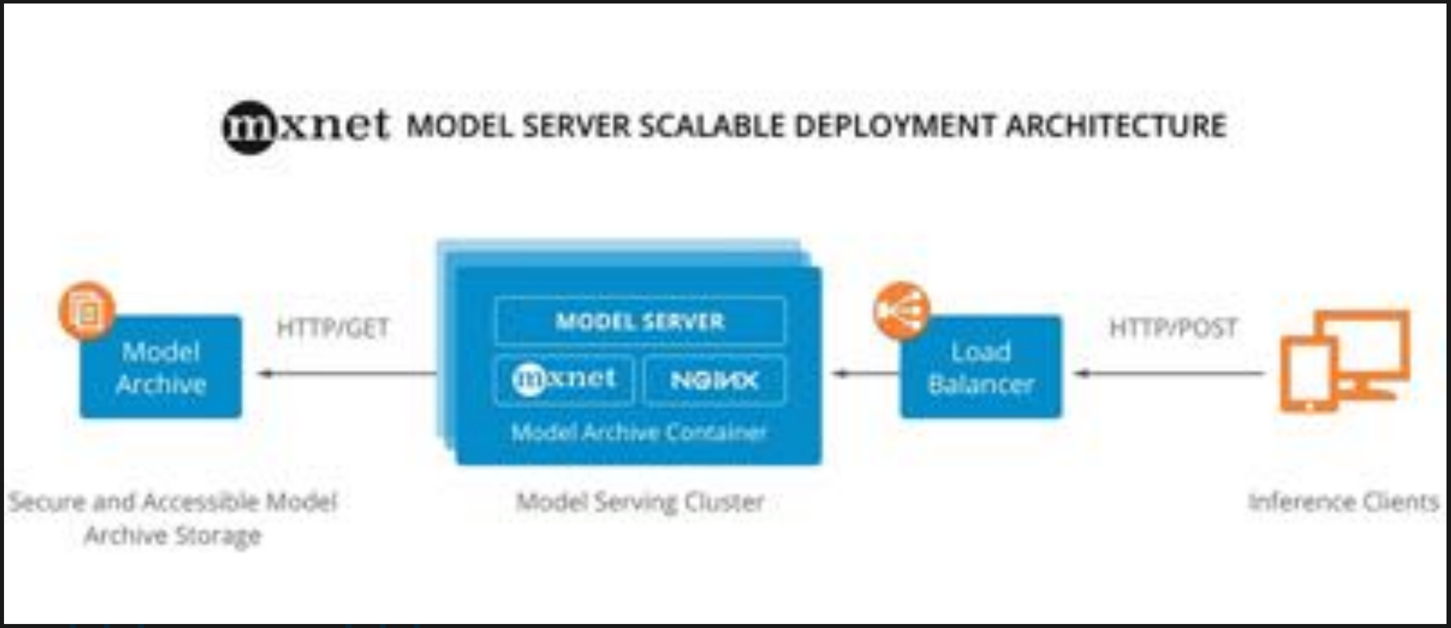
public > VisualSearch_MXNet

Name	Last Modified
images	an hour ago
visualsearch.ipynb	an hour ago
README.md	an hour ago
skirt.jpg	an hour ago

```
plot_predictions(images)
```

Model Serving

MXNet Model Server Architecture



Back

Amazon SageMaker Neo – a path to IoT

- Model compiler supporting various hardware platforms incl. Intel, Nvidia, Arm, Cadence, Qualcomm, and Xilinx
- Supporting Tensorflow, MXNet and ONNX
- https://tvm.ai/

Thank you!

Connect here
github.com/srochel
linkedin.com/in/steffenrochel
steroche@amazon.com

mxnet-info@amazon.com

Photos credits: https://pexels.com and https://unsplash.com/