Powered by Plain Concepts

#STechDay2021

SINGULARITY

TECH DAY_2021

The era of AI and Cognitive Services

¿Es este el final de las redes neuronales convolucionales?

plain concepts

databricks

ORGANIZATION

SPONSORS

THANK YOU!

Rodrigo Cabello

Research Engineer Microsoft® MVP Most Valuable Professional

Research Engineer at Plain Concepts and Microsoft MVP in Artificial Intelligence with more than 10 years of experience in software development. Passionate about cutting-edge technologies and everything related to Artificial Intelligence. Currently, I develop state-of-the-art machine learning solutions, mainly in the area of Computer Vision.

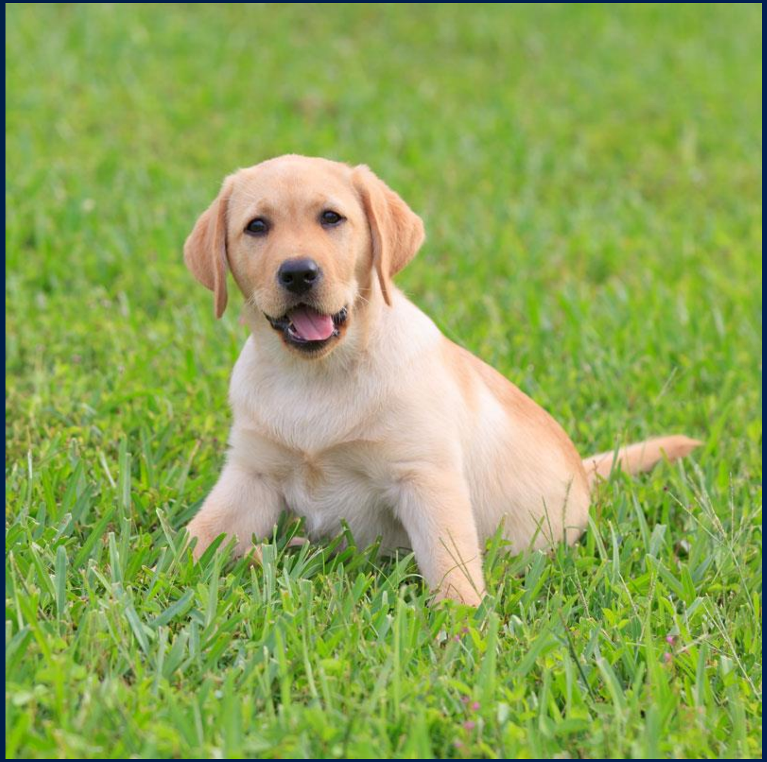
@mrcabellom

✉ mrcabello@plainconcepts.com

“Convolutional neural networks lacks a global understanding of the images. It only looks for the presence of the image's features and does not understand the structural dependency between its features.”

Capsule networks 2017

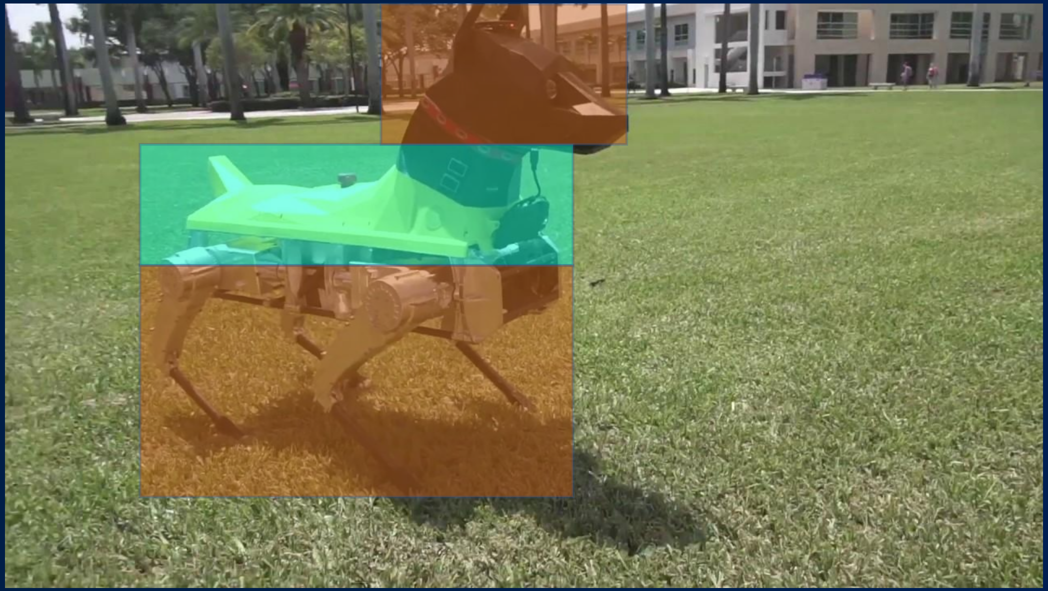
Computer vision



```
"captions": [
        {
          "text": "a large brown dog lying on green grass",
          "confidence": 0.8984215667012988
        }
    ]

   "objects": [
        {
          "object": "golden retriever",
          "confidence": 0.609
        }
    ]
```

Computer vision

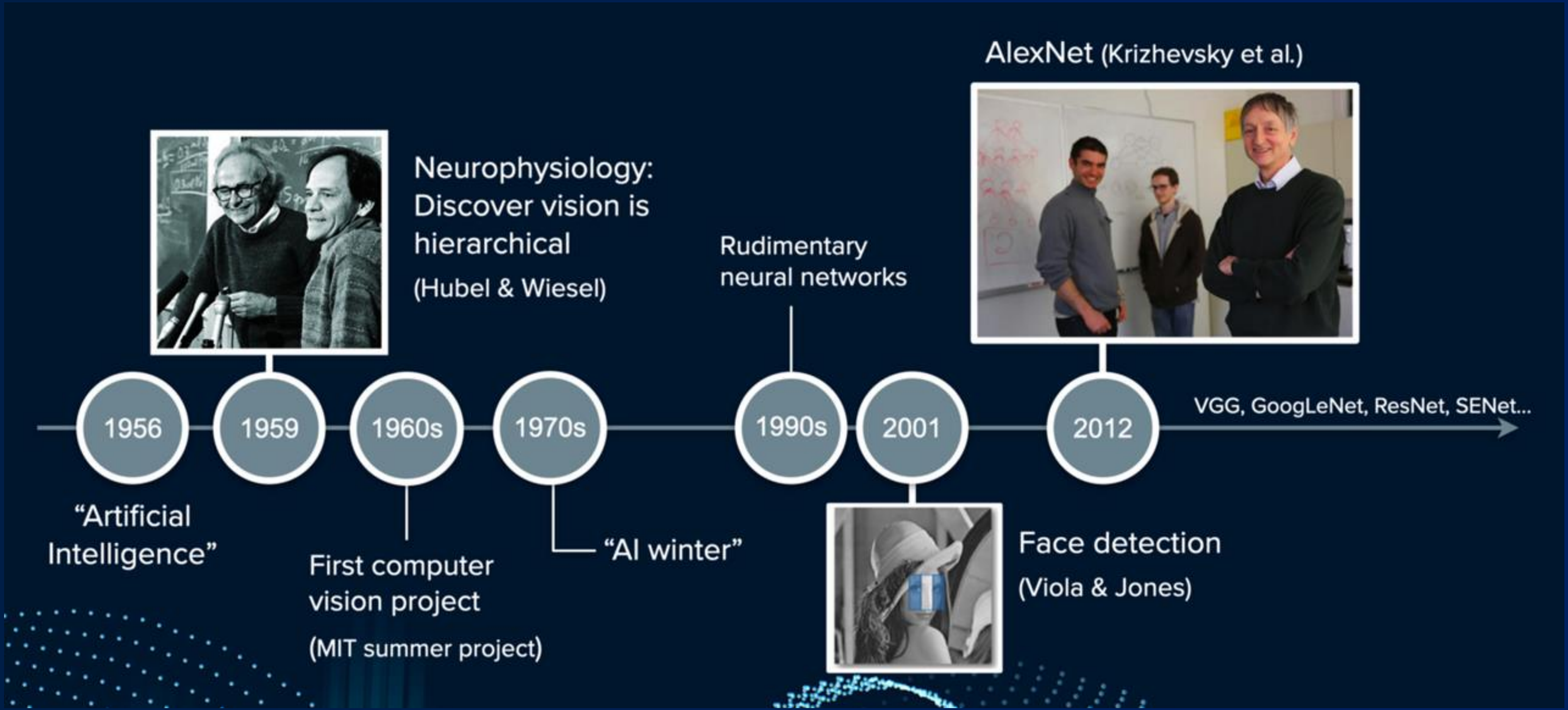


```
"captions": [
        {
            "text": "a dog sitting on top of a grass covered field",
            "confidence": 0.7082839064777712
        }
    ]

  "objects": [
        {
              "object": "aircraft",
            "confidence": 0.57,
        }
  ]
```

Computer vision

Computer vision timeline



1956 "Artificial Intelligence"

1959 Neurophysiology: Discover vision is hierarchical

(Hubel & Wiesel)

1960s First computer vision project

(MIT summer project)

1970s "AI winter"

1990s Rudimentary neural networks

2001 Face detection

(Viola & Jones)

2012 AlexNet (Krizhevsky et al.)

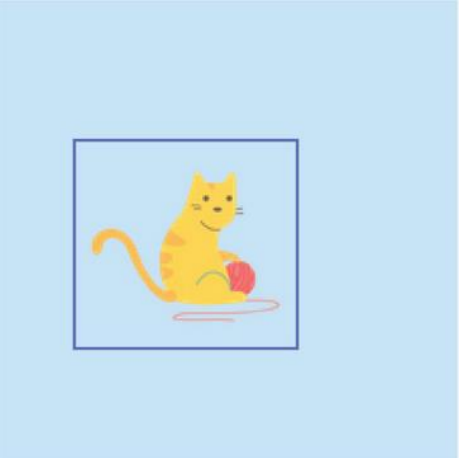
VGG, GoogLeNet, ResNet, SENet...

Computer vision use cases

Classification

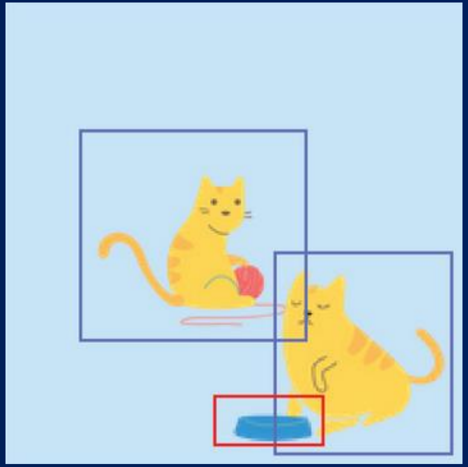
CAT

Classification + Localization



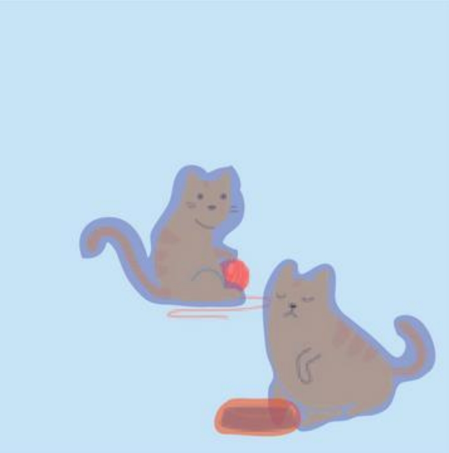
CAT + Bounding Box

Object detection



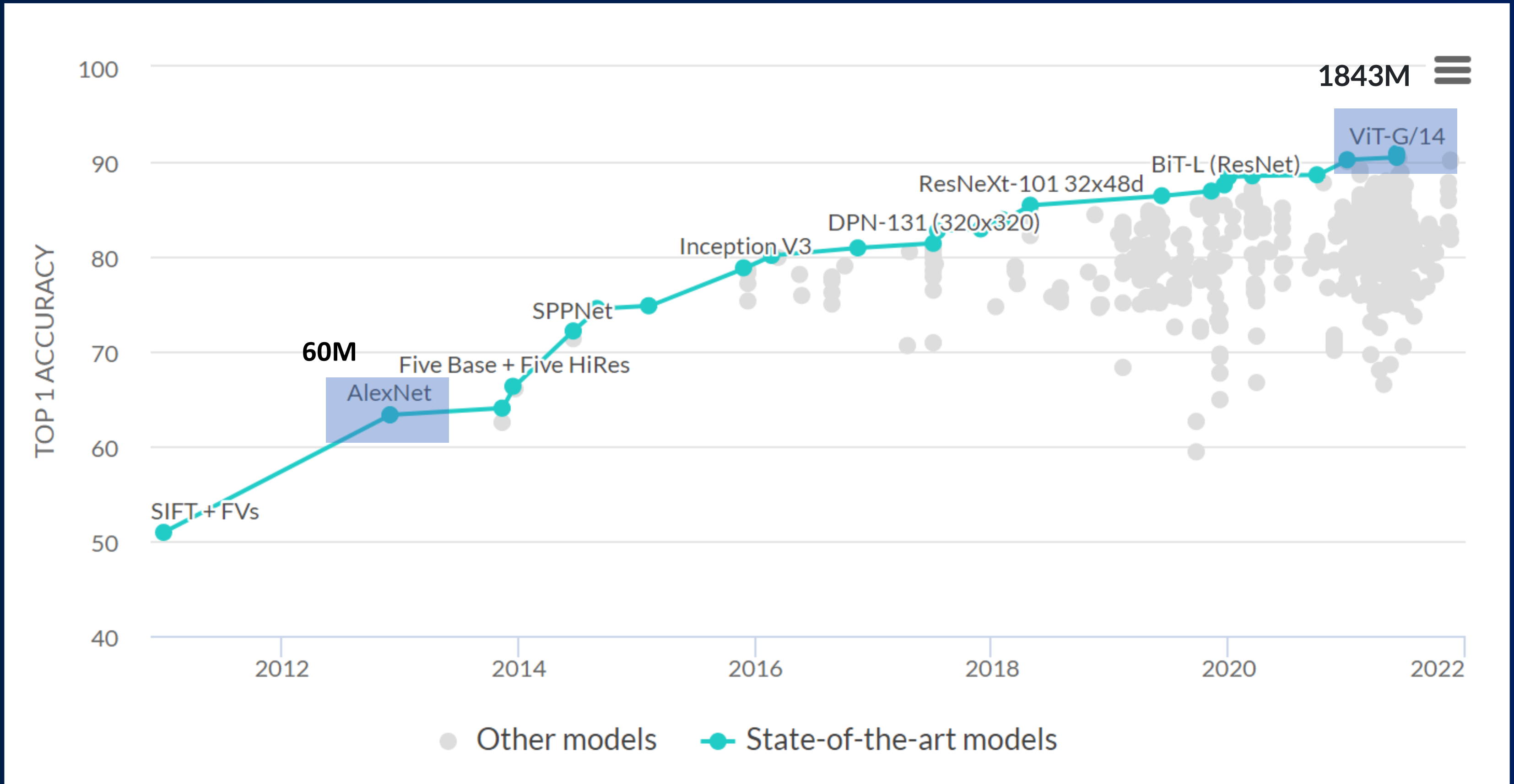
CAT, CAT,BOWL

Semantic segmentation



CAT, CAT, BOWL

Computer vision neural networks



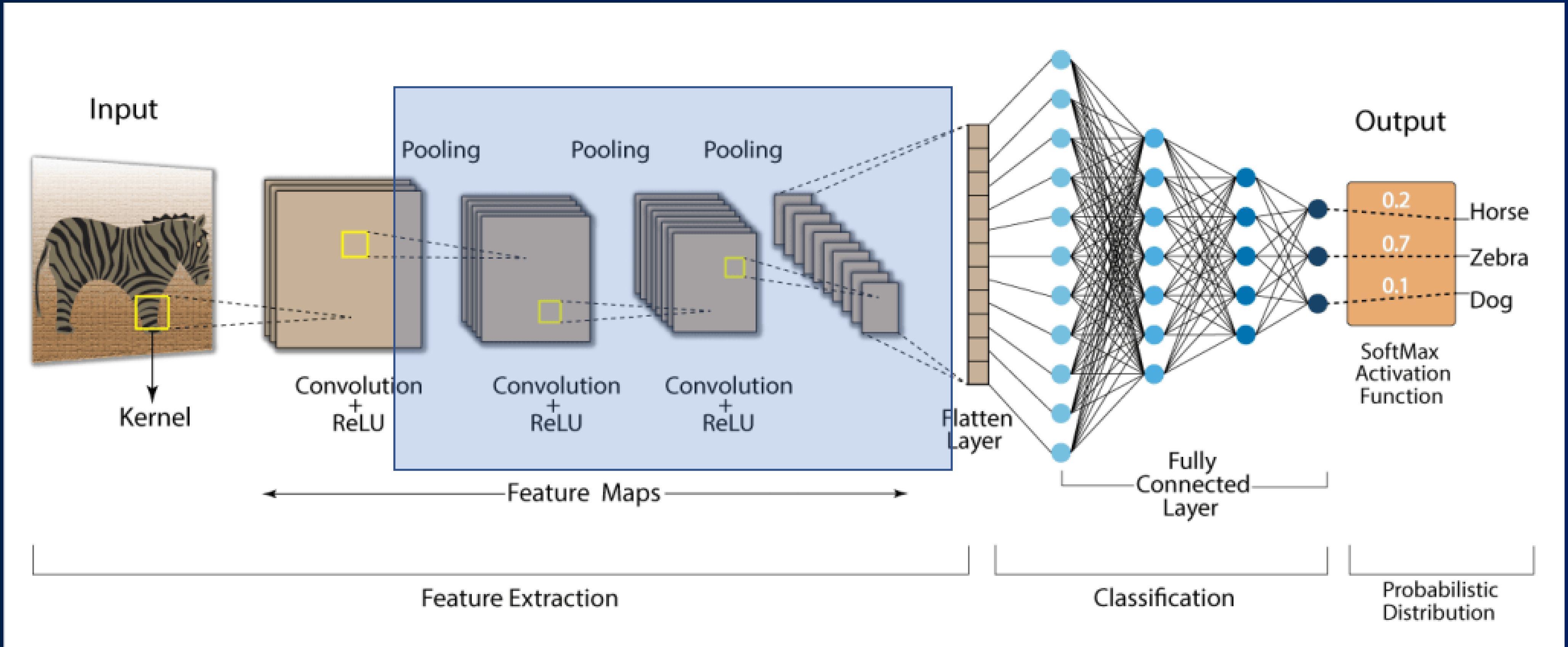
ImageNet Benchmark

Vision Transformers
2020

Capsule networks??
2017

Convolutional neural networks

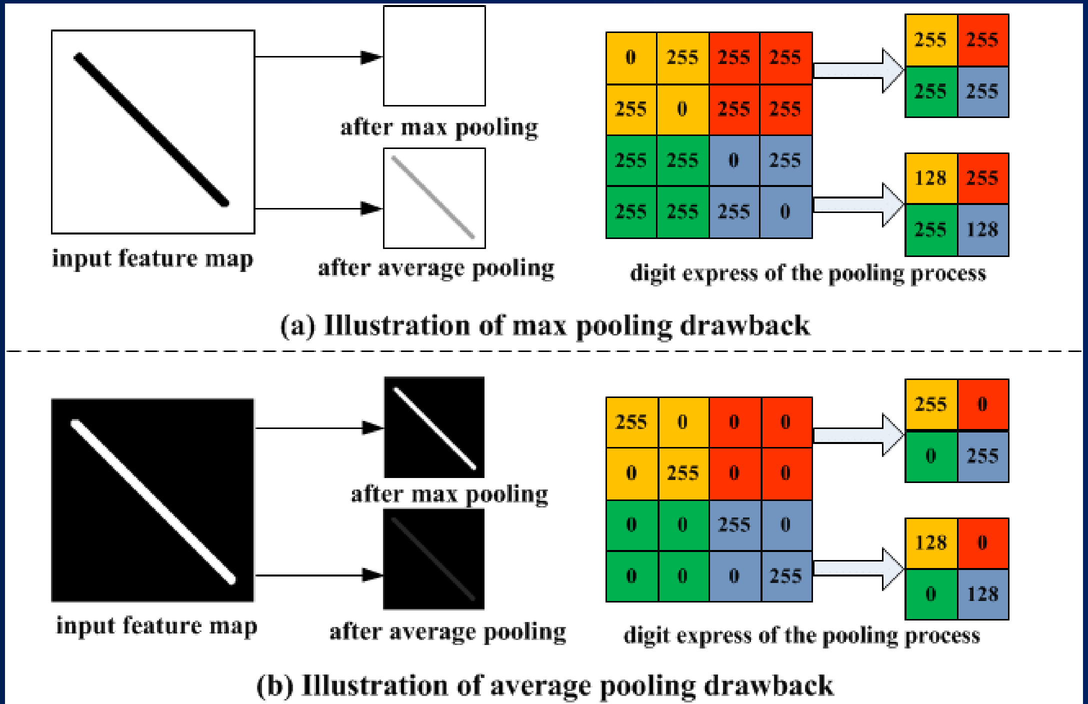
CNN topology



Problems with CNNs

- CNNs detect certain features in an image, but through using a pooling layer valuable information gets lost.

- CNN use “pooling” or equivalent methods to “summarize” what's going on in the smaller regions and make sense of larger and larger chunks of the image. This was a solution that made CNNs work well, but it loses valuable information.



0	255	255	255
255	0	255	255
255	255	0	255
255	255	255	0

255	255
255	255

128	255
255	128

(a) Illustration of max pooling drawback

255	0	0	0
0	255	0	0
0	0	255	0
0	0	0	255

255	0
0	255

128	0
0	128

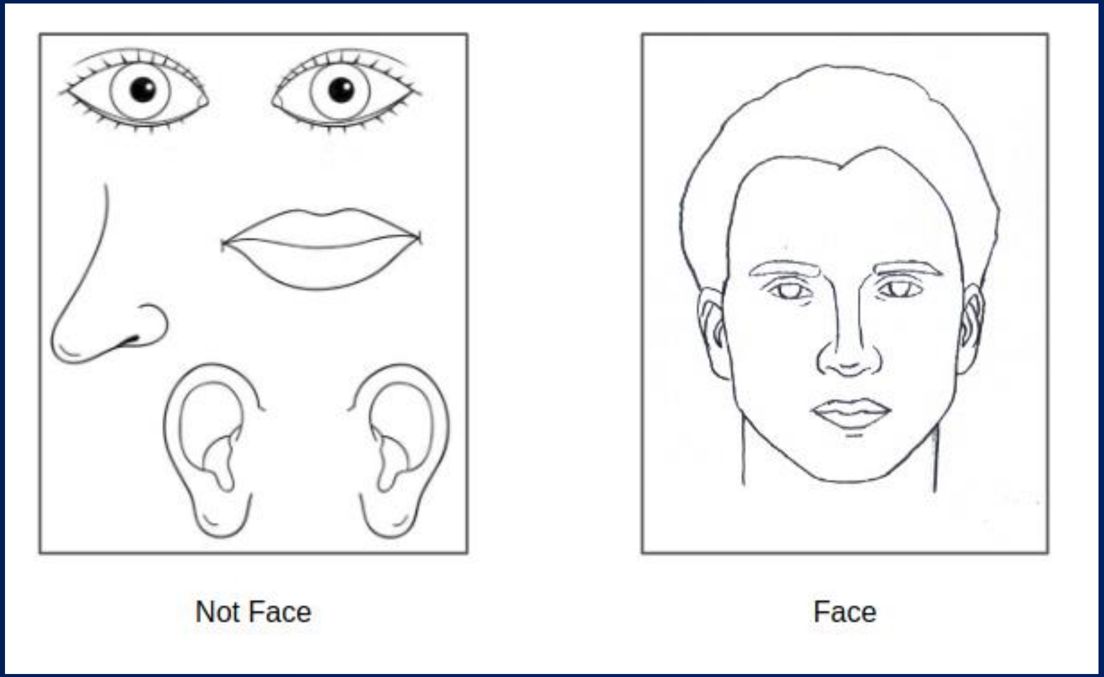
(b) Illustration of average pooling drawback

“The pooling operation used in convolutional neural networks is a big mistake, and the fact that it works so well is a disaster.”

Geoffrey Hinton

Problems with CNNs

- **Is not scale and rotation invariant.**
  - Data augmentation
- **Translation invariance**
  - Ignores the relation between the part and the whole.
- **Receptive field**
  - Model long range dependencies



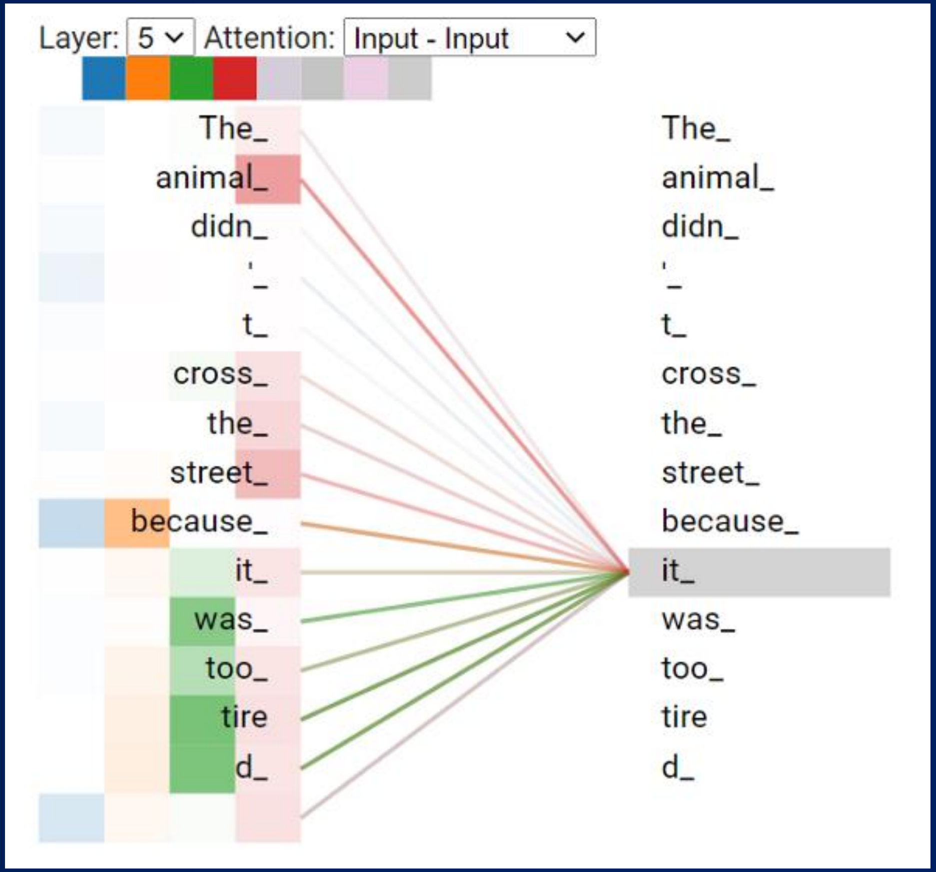
Vision Transformers

Transformers

- It is used primarily in the field of natural language processing (NLP).
- Transformers are designed to handle sequential input data. The attention mechanism provides context for any position in the input sequence.
- GPT-3 has a transformers-based architecture. It solves tasks such as:
  - Translation, text summarization, semantic search, questions answering, document classification

Transformers (self-attention)

- Attention are trainable weights that model the importance of each part of an input sentence.
- It will look at each word of the sentence and compare its position in the sentence with respect to the position of all the words present in the same sentence (including itself).
- A score is calculated based on these positional clues which is then used to encode the semantics or meaning of the sentence in a better way.



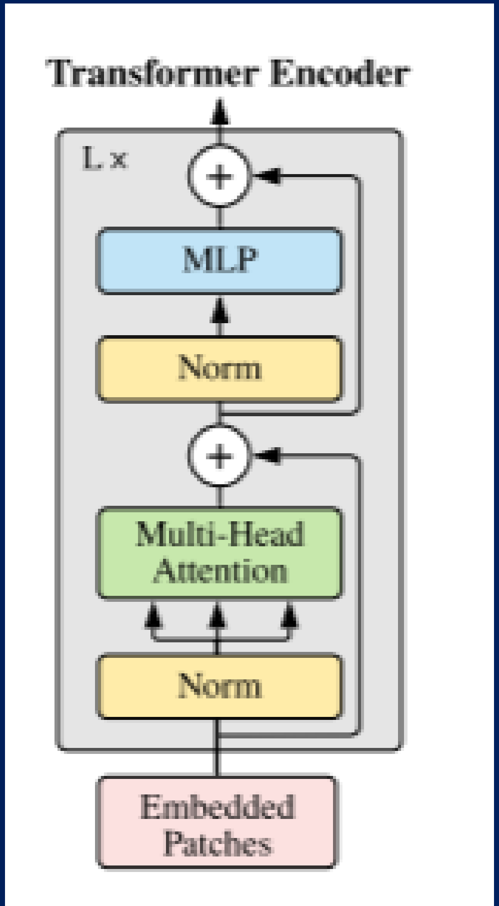
Layer: 5 Attention: Input - Input

Vision transformers

- Vision transformer uses pixels to achieve a similar result for images.
- The image is divided into small patches. After that, all patches are flattened using a linear projection.
- A classification head is attached at the end of the transformer encoder to predict the final classes.

Transformer encoder

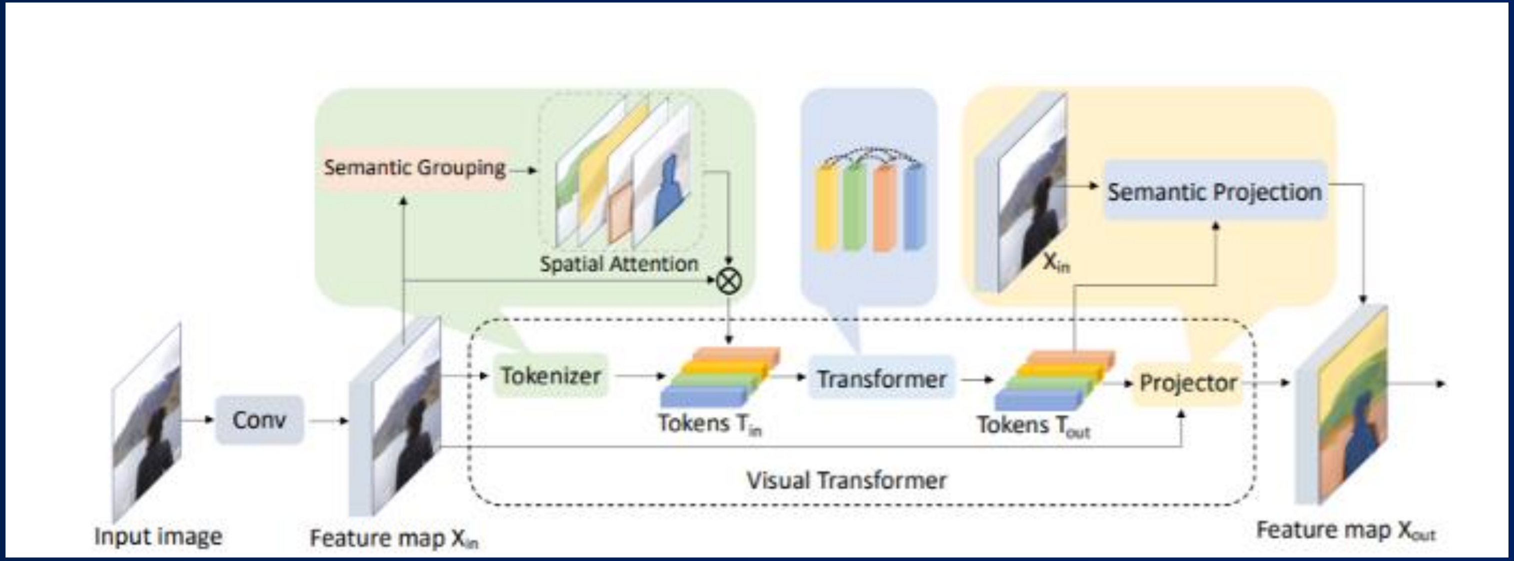
- **Multi-Head Self Attention Layer(MSP)** to concatenate the multiple attention outputs linearly to expected dimensions.
- **Multi-Layer Perceptrons(MLP)** contains two-layer with Gaussian Error Linear Unit(GELU).
- **Layer Norm(LN)** is applied before every block as it does not introduce any new dependencies between the training images
- **Residual connections** are applied after every block as they allow the gradients to flow through the network directly without passing through non-linear activations.



Transformer architectures

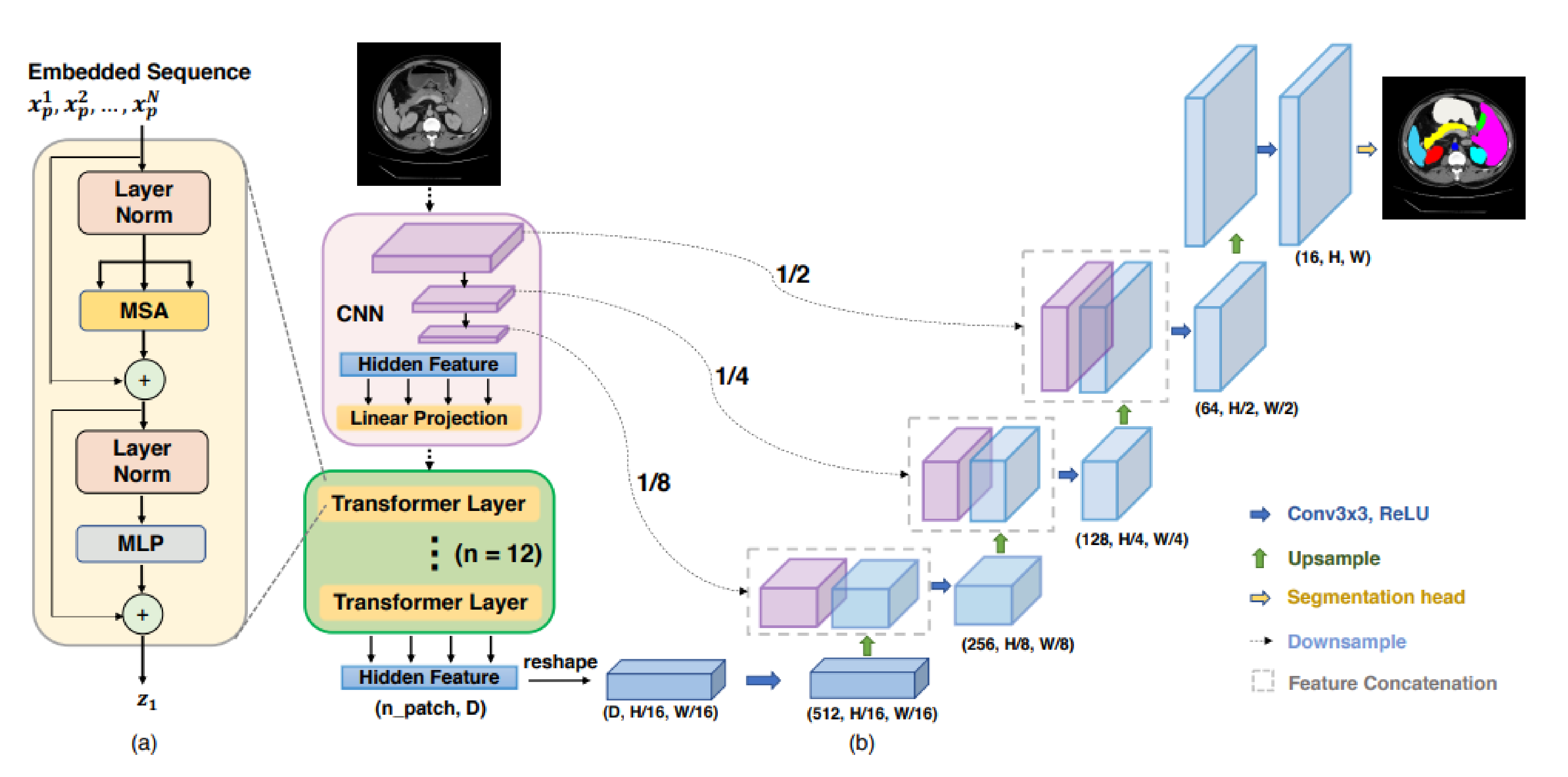
Hybrid Architectures

- Use CNN as feature extraction.
- Input sequence based on feature maps, then followed by applying the encoding to the feature patches.



TransUnet

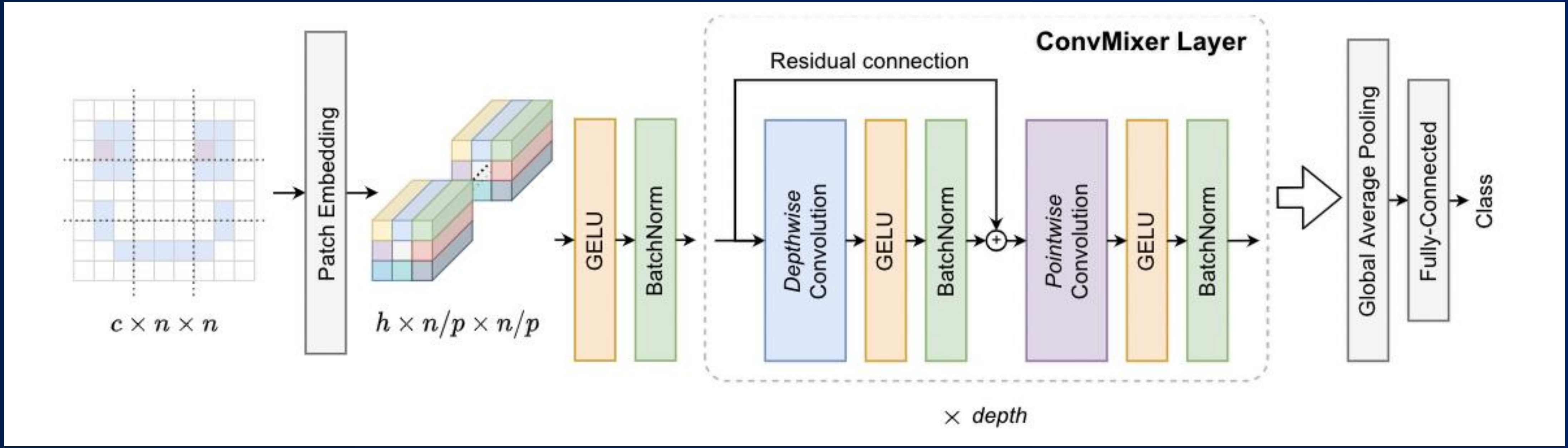
- CNN-Transformer Hybrid as Encoder.
- Cascaded Upsampler.



Demo

What is next?

ConvMixers: Patches are all you need?

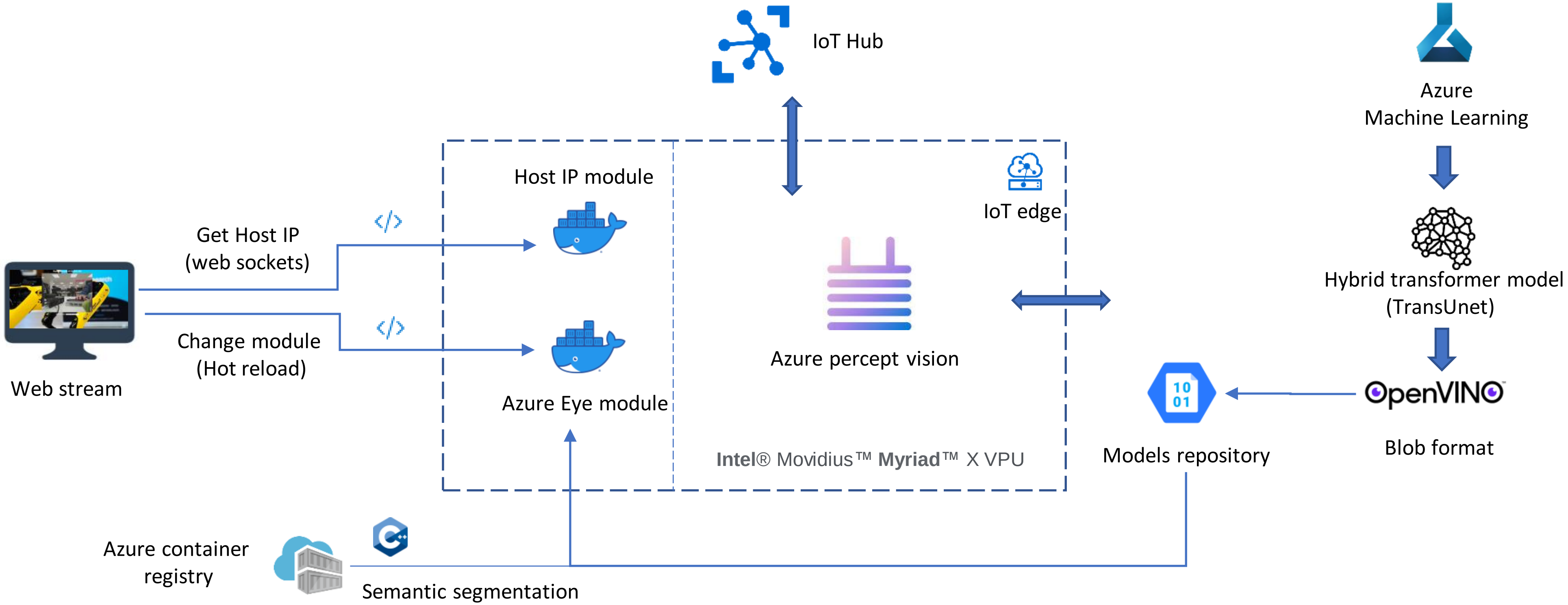


Patch representation

Isotropically repeated convolutional blocks.

Demo

Architecture



[Webstream Video (windows.net)]

SINGULARITY
TECH DAY_2021

Thank you!

databricks

plain concepts