MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

Universidad de Valparaíso CHILE

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Explainable AI (XAI) for Image Analysis with Deep Learning

Dr. Rodrigo Salas Fuentes
rodrigo.salas@uv.cl

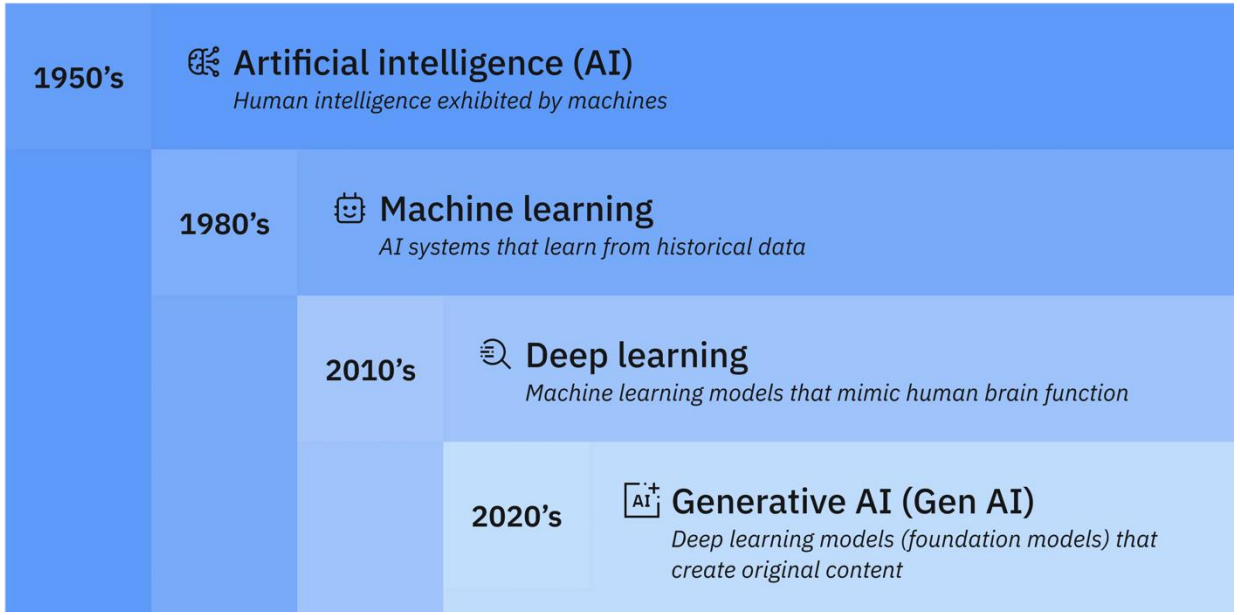
MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Artificial Intelligence
AI

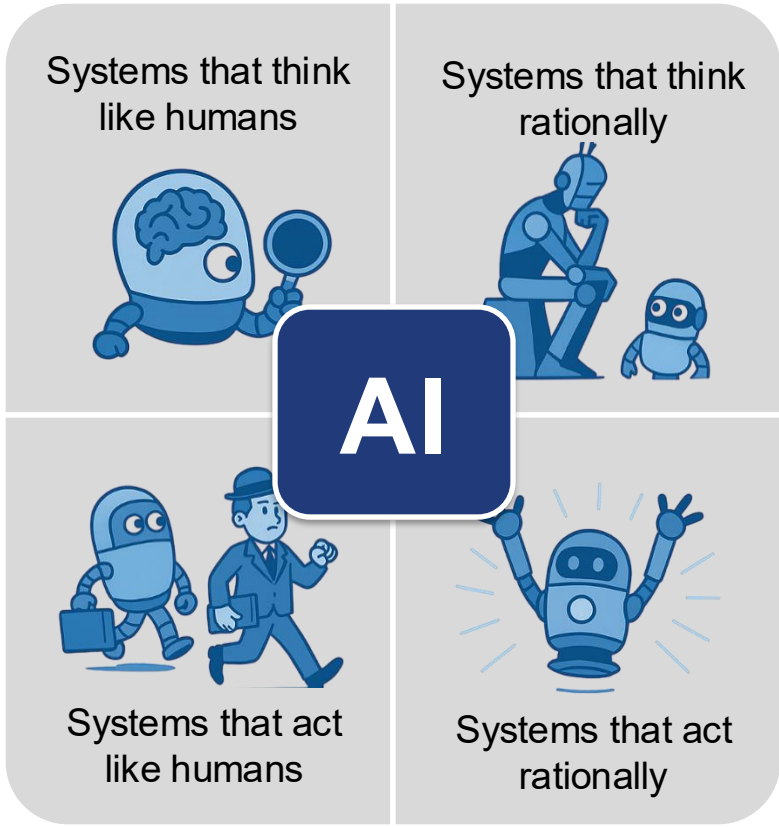
Universidad de Valparaíso CHILE

What is Artificial Intelligence



https://www.ibm.com/mx-es/think/topics/sentient-ai

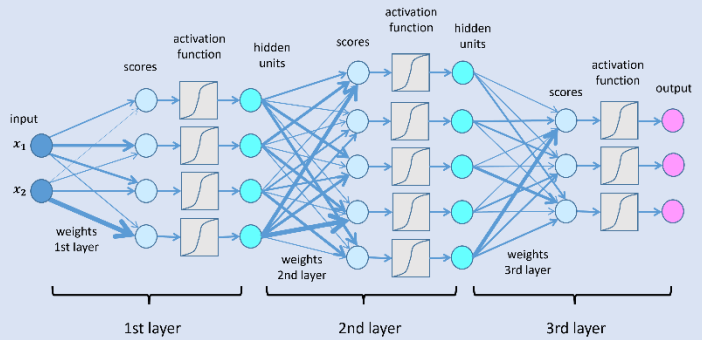
Artificial Intelligence (AI) is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy.



Stuart Russell y Peter Norvig, 2020

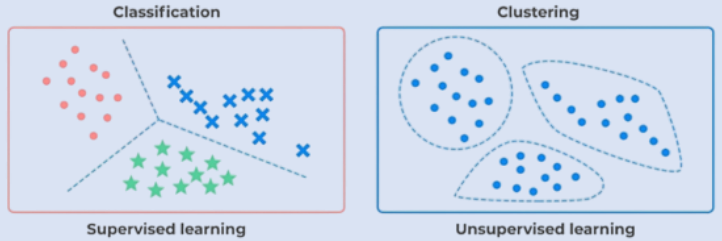
AI Methods

https://lamarr-institute.org/blog/deep-neural-networks/



Deep Learning

https://www.superannotate.com/blog/image-classification-basics

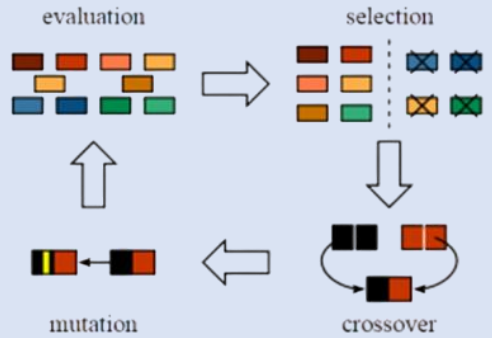


Machine Learning

https://devopedia.org/natural-language-processing

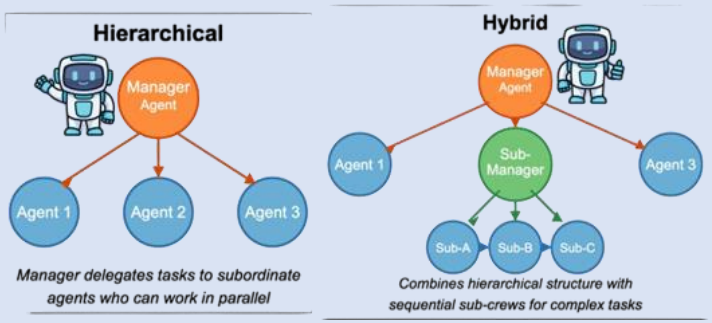
Natural Language Processing

https://doi.org/10.1145/3446132.3446142



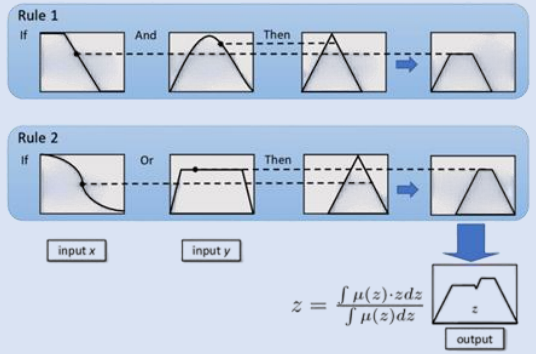
Metaheuristics

https://l1nq.com/at8FT



Multi-Agent Systems

http://dx.doi.org/10.1016/j.ins.2018.09.005



Fuzzy Inference Systems

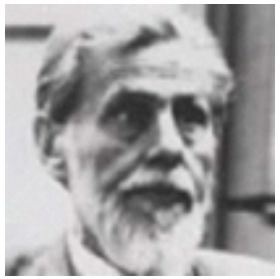
MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

ihealth | Millennium Institute for Intelligent Healthcare Engineering

Artificial Neural Networks
ANN

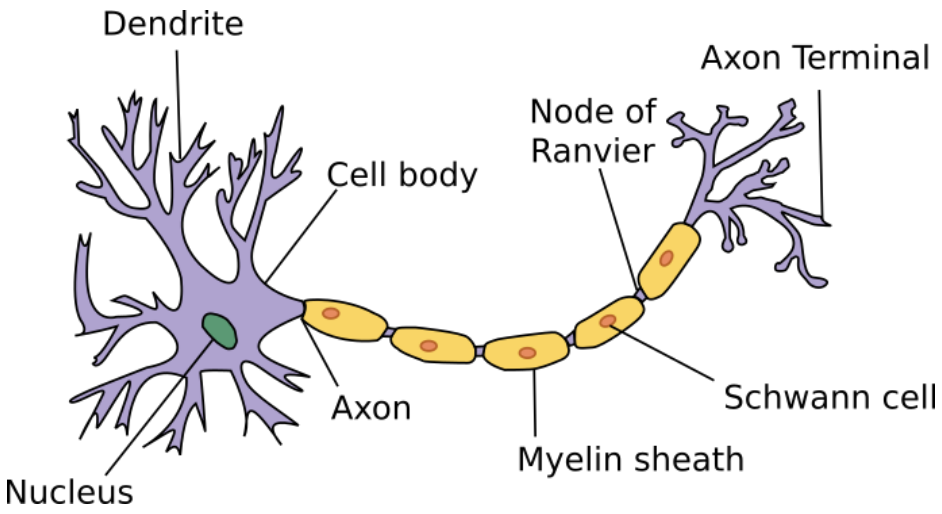
Universidad de Valparaíso CHILE

The Artificial Neuron

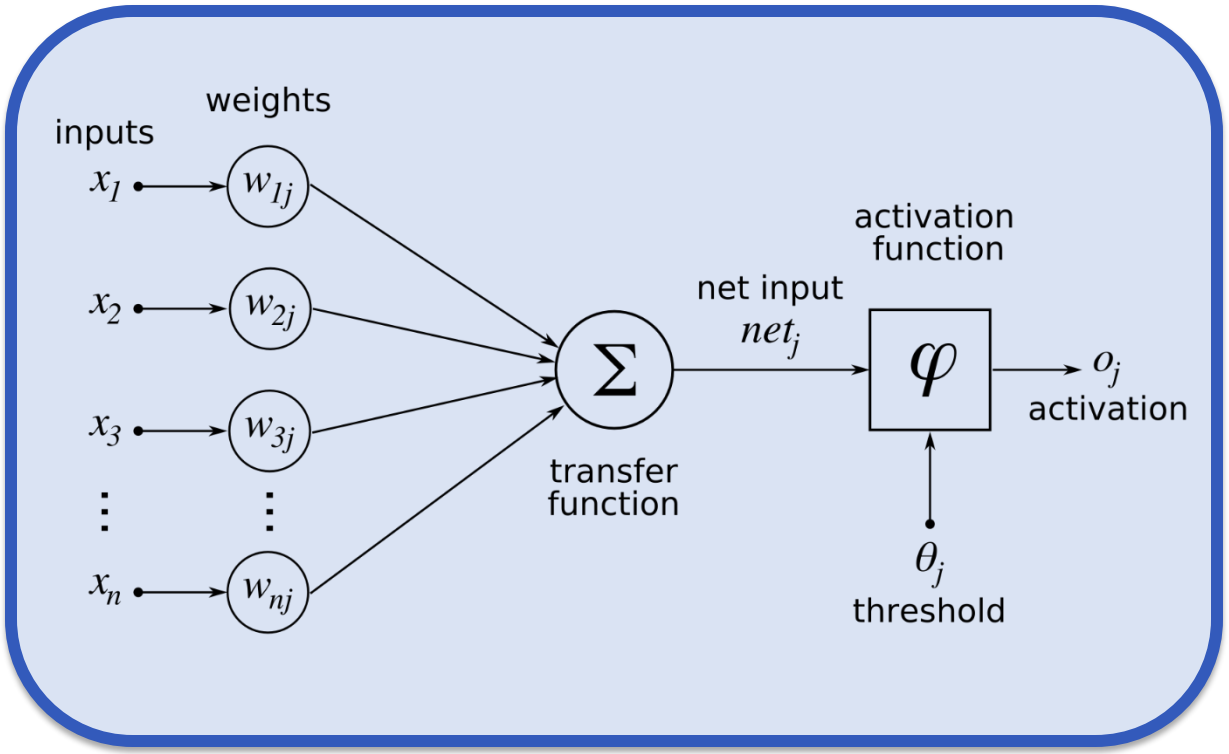


Warren McCulloch (1898 - 1972)

Walter Pitts (1924 - 1969)



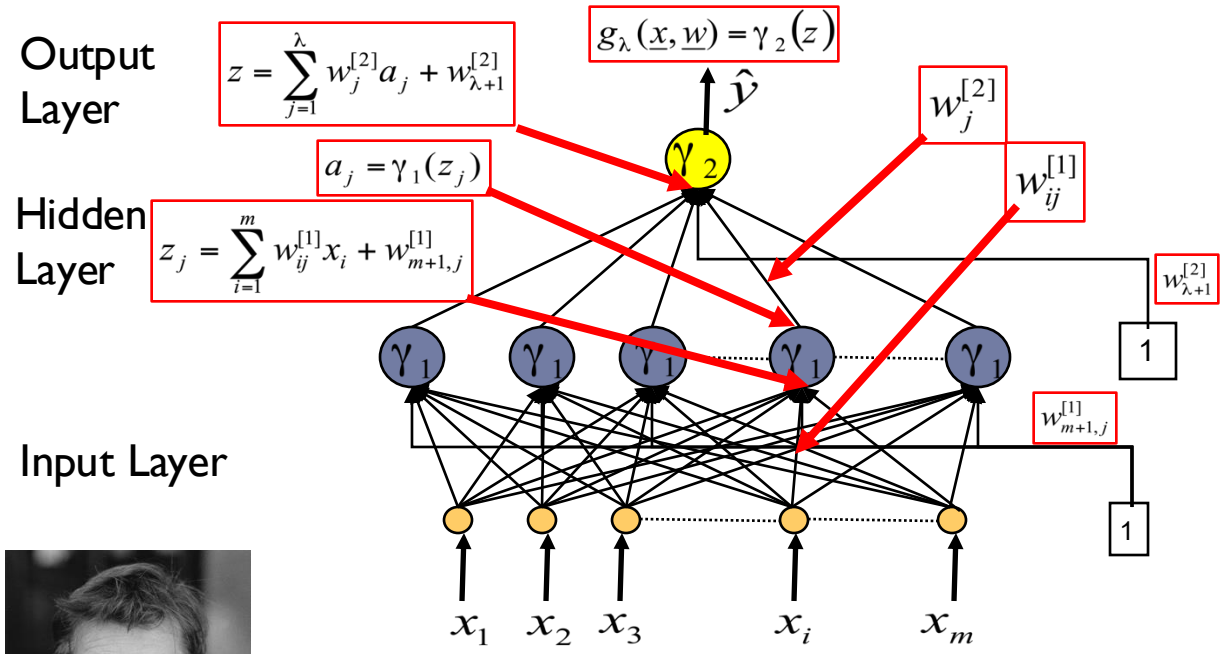
Warren McCulloch and Walter Pitts, A Logical Calculus of Ideas Immanent in Nervous Activity, 1943, Bulletin of Mathematical Biophysics 5:115-133.

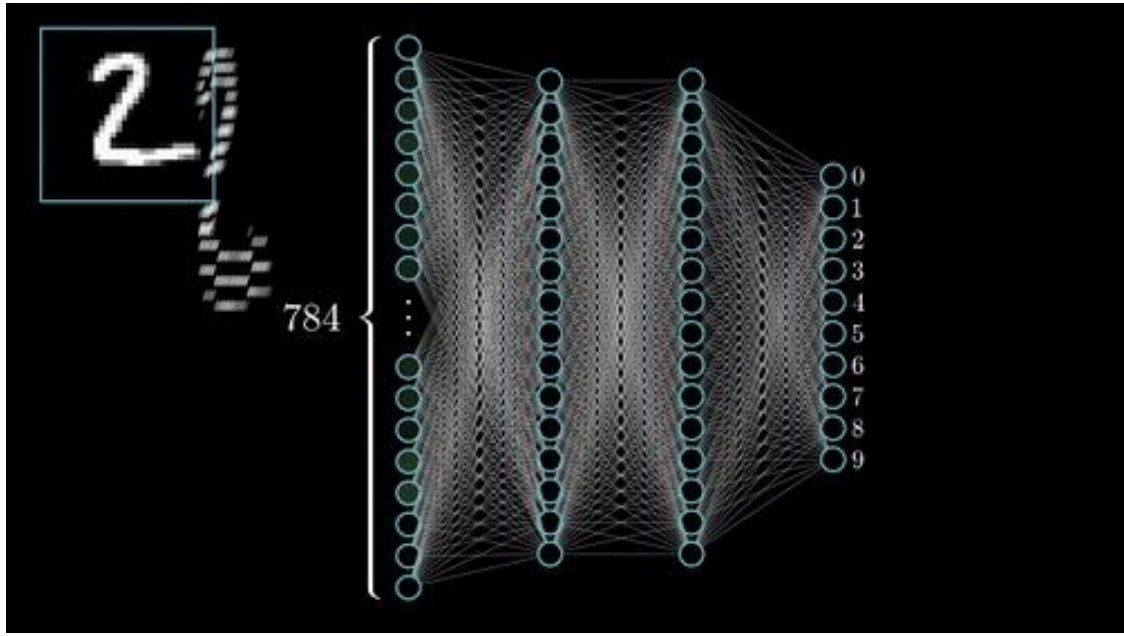


https://afit-r.github.io/ann_fundamentals

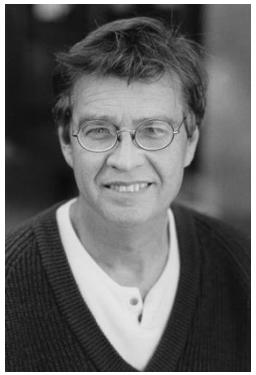
Multilayer Perceptron (MLP)

$$g_\lambda(\underline{x},\underline{w}) = \gamma_2\left(\sum_{j=1}^{\lambda} w_j^{[2]}\gamma_1\left(\sum_{i=1}^{m} w_{ij}^{[1]}x_i + w_{m+1,j}^{[1]}\right) + w_{\lambda+1}^{[2]}\right)$$



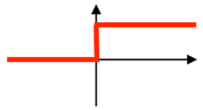
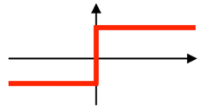
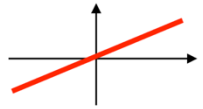
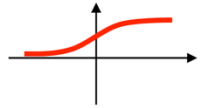
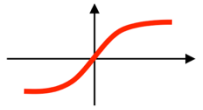
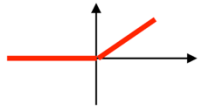
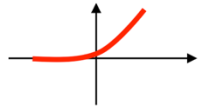


https://medium.com/@Suraj_Yadav/in-depth-knowledge-of-convolutional-neural-networks-b4bfff8145ab



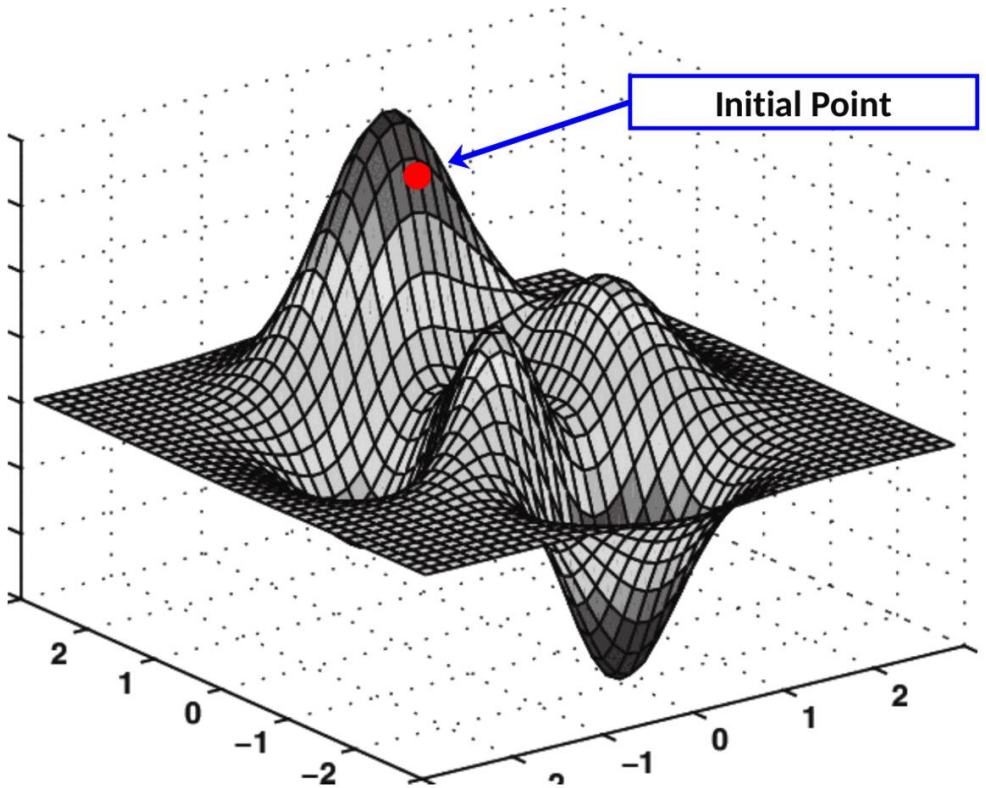
Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation", Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986

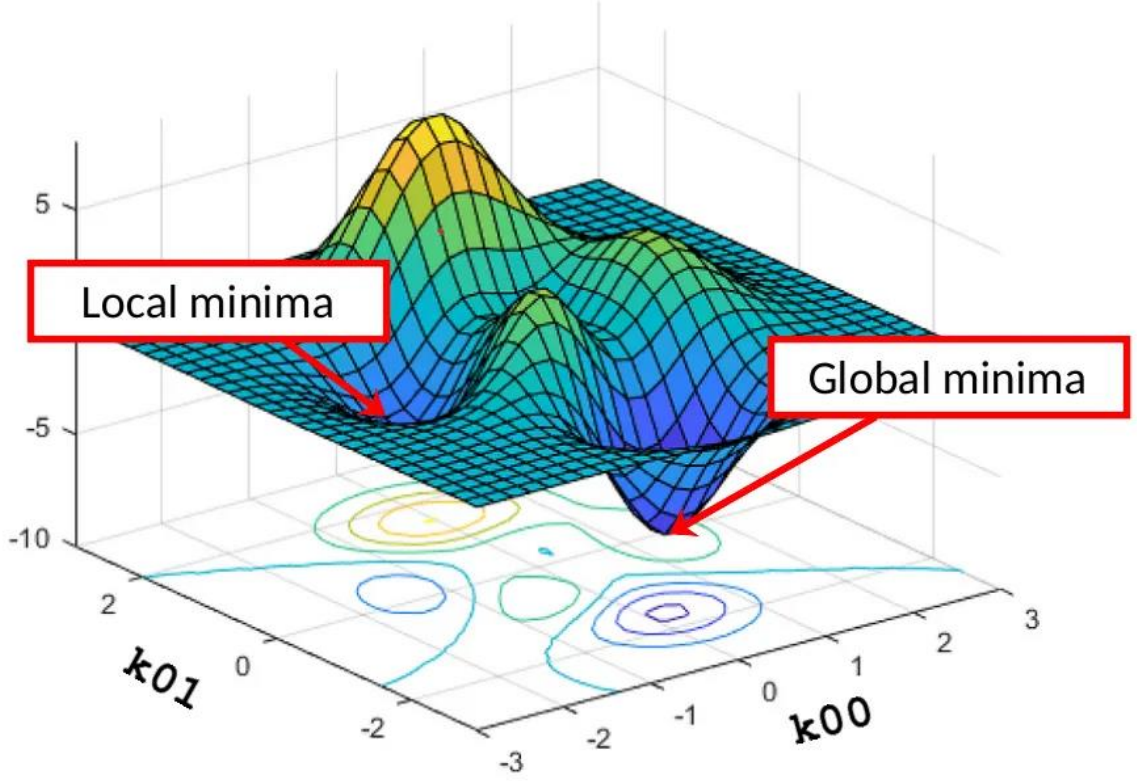
Activation Functions

Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^{z})$	Multi-layer Neural Networks	

Delta Rule

$$w_j(t+1) = w_j(t) + \alpha(y - \hat{y})x_j$$



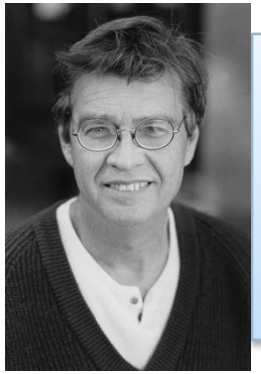


https://pub.towardsai.net/deep-learning-from-scratch-in-modern-c-gradient-descent-670bc5889112

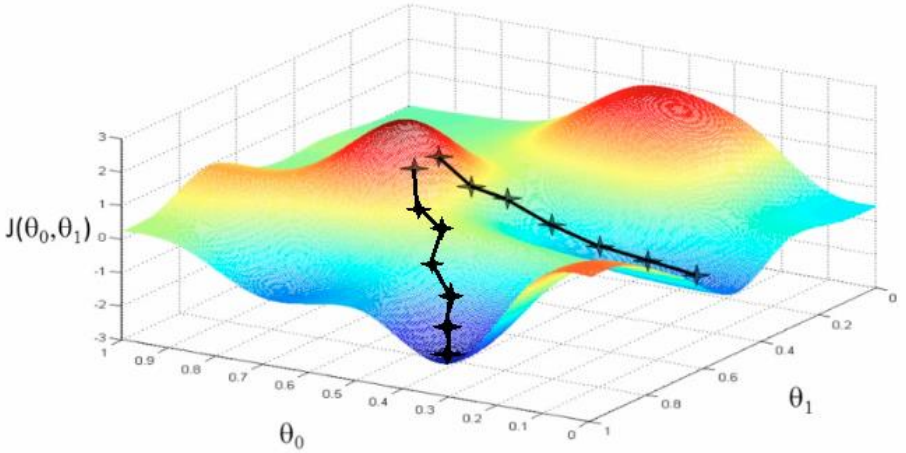
Backpropagation

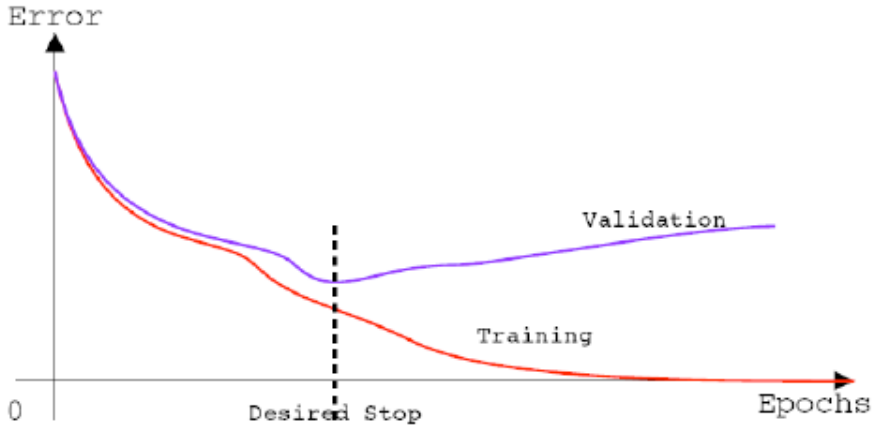
$$\mathbf{w}(t+1) = \mathbf{w}(t) - \alpha(t)\nabla_{\mathbf{w}} l(h(\cdot, \mathbf{w}); z_t)$$

Paul Werbos.
1974 | In his Harvard PhD thesis, Paul Werbos describes training neural networks through backpropagation.



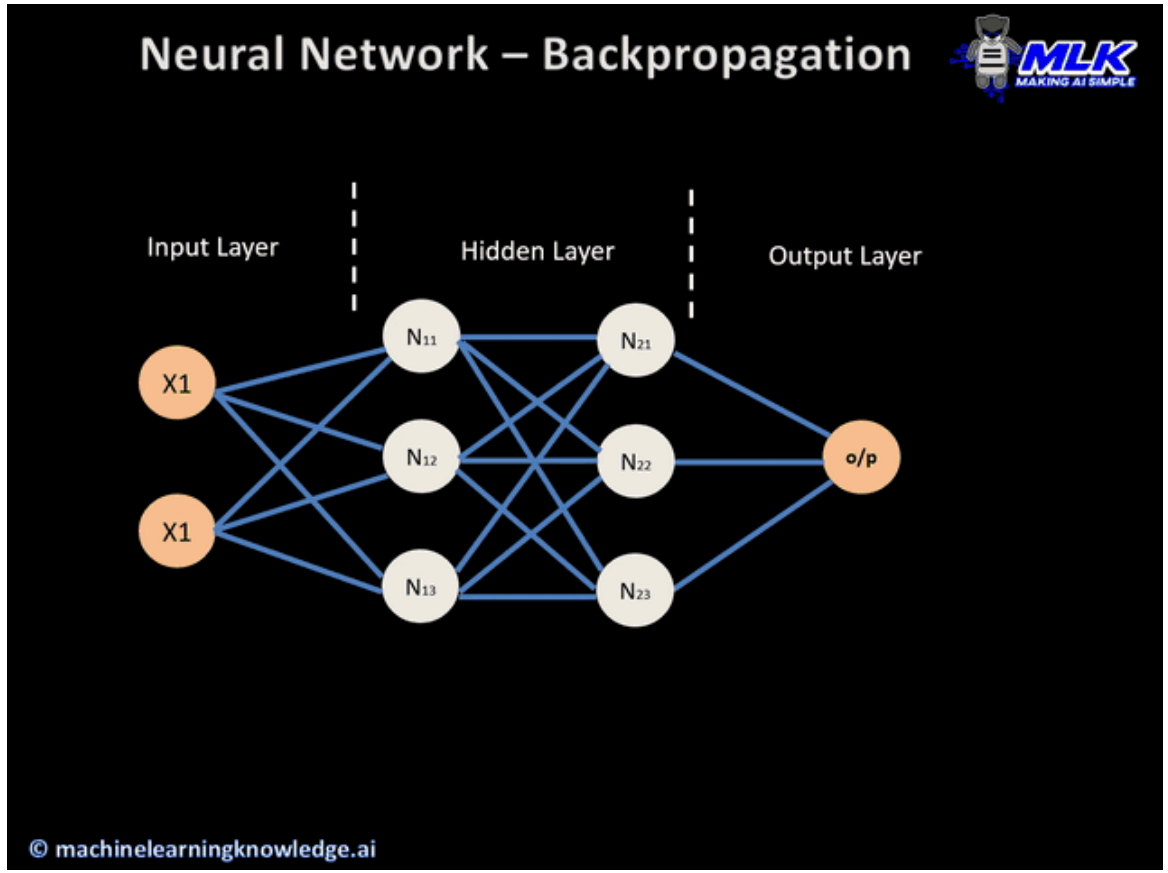
Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation", Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986

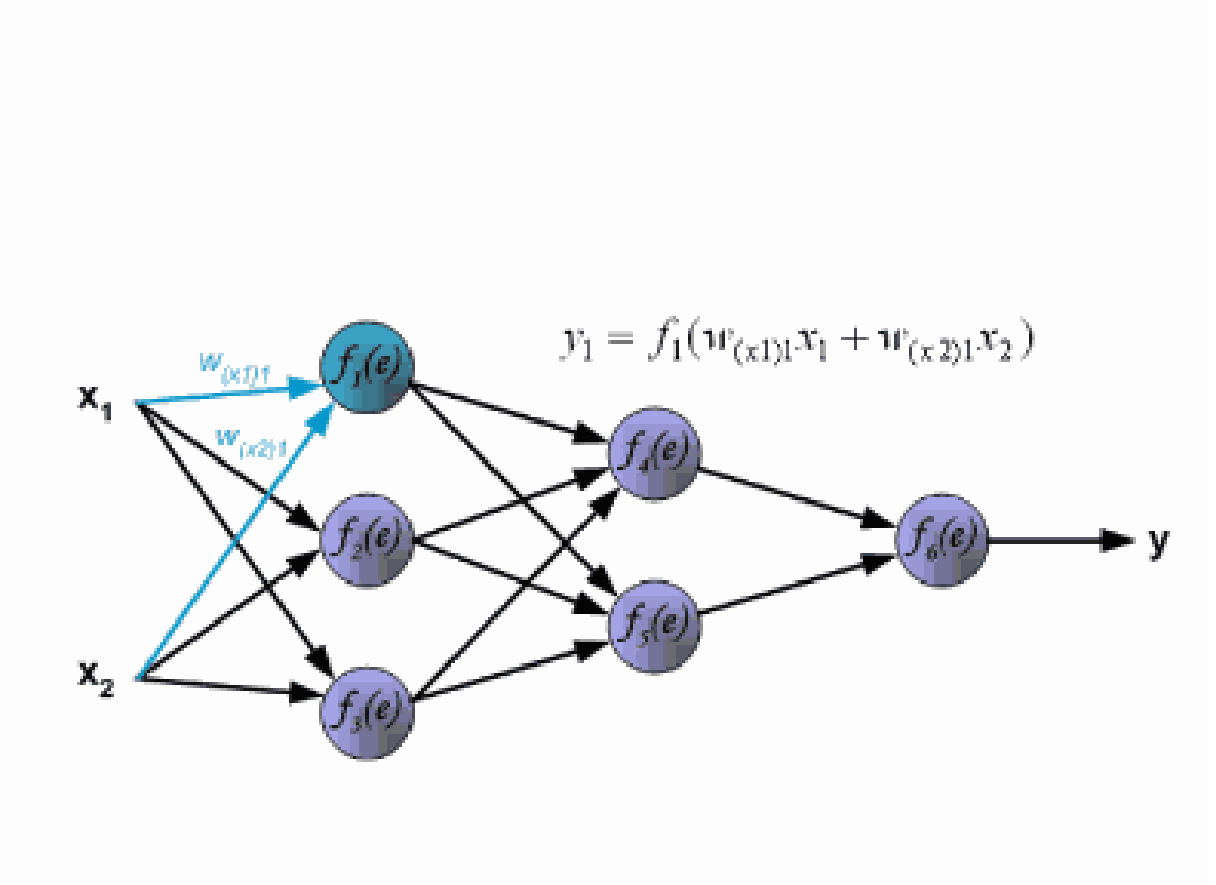




http://www.holehouse.org/mlclass/01_02_Introduction_regression_analysis_and_gr.html

Backpropagation





https://medium.com/deeper-deep-learning-tr/adım-adım-forward-and-back-propagation-cf4cd18276ee

Backpropagation Algorithm

$$\frac{\partial L}{\partial w_{(h_{r-1},h_r)}} = \delta(h_r, o) \cdot h_{r-1}$$

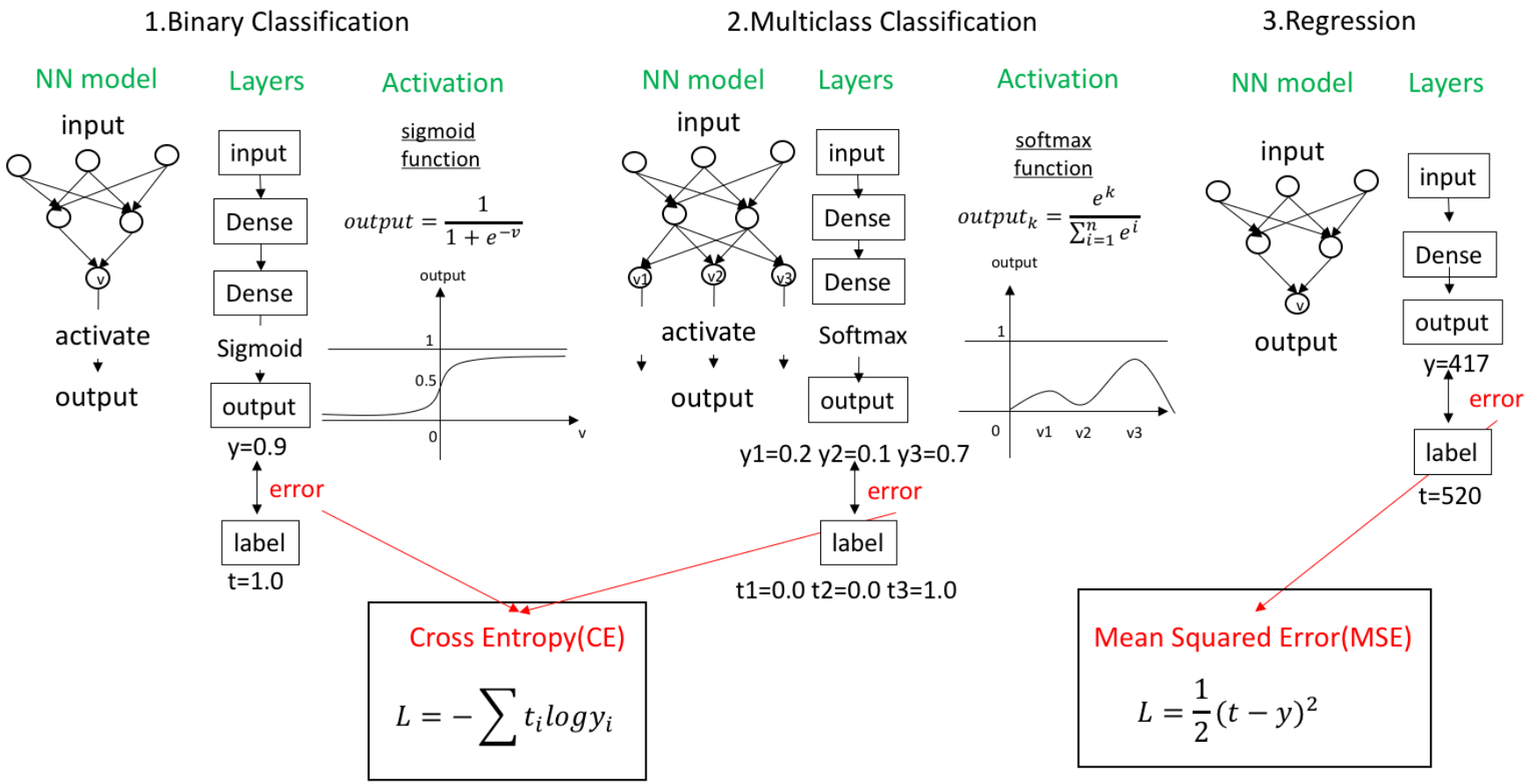
$$\delta(o,o) = \frac{\partial L}{\partial a_o} = \Phi'(a_o) \cdot \frac{\partial L}{\partial o}$$

$$\frac{\partial L}{\partial w_{(h_{r-1},h_r)}} = \frac{\partial L}{\partial o} \cdot \Phi'(a_o) \cdot \left[\sum_{[h_r, h_{r+1}, \ldots h_k, o] \in \mathcal{P}} \frac{\partial a_o}{\partial a_{h_k}} \prod_{i=r}^{k-1} \frac{\partial a_{h_{i+1}}}{\partial a_{h_i}} \right] \underbrace{\frac{\partial a_{h_r}}{\partial w_{(h_{r-1},h_r)}}}_{h_{r-1}}$$

Backpropagation computes $\delta(h_r, o) = \frac{\partial L}{\partial a_{h_r}}$

$$\delta(h_r,o) = \frac{\partial L}{\partial a_{h_r}} = \sum_{h:h_r \Rightarrow h} \overbrace{\frac{\partial L}{\partial a_h}}^{\delta(h,o)} \underbrace{\frac{\partial a_h}{\partial a_{h_r}}}_{\Phi'(a_{h_r}) w_{(h_r,h)}} = \Phi'(a_{h_r}) \sum_{h:h_r \Rightarrow h} w_{(h_r,h)} \cdot \delta(h,o)$$

Loss Functions



https://www.renom.jp/notebooks/tutorial/basic_algorithm/lossfunction/notebook.html

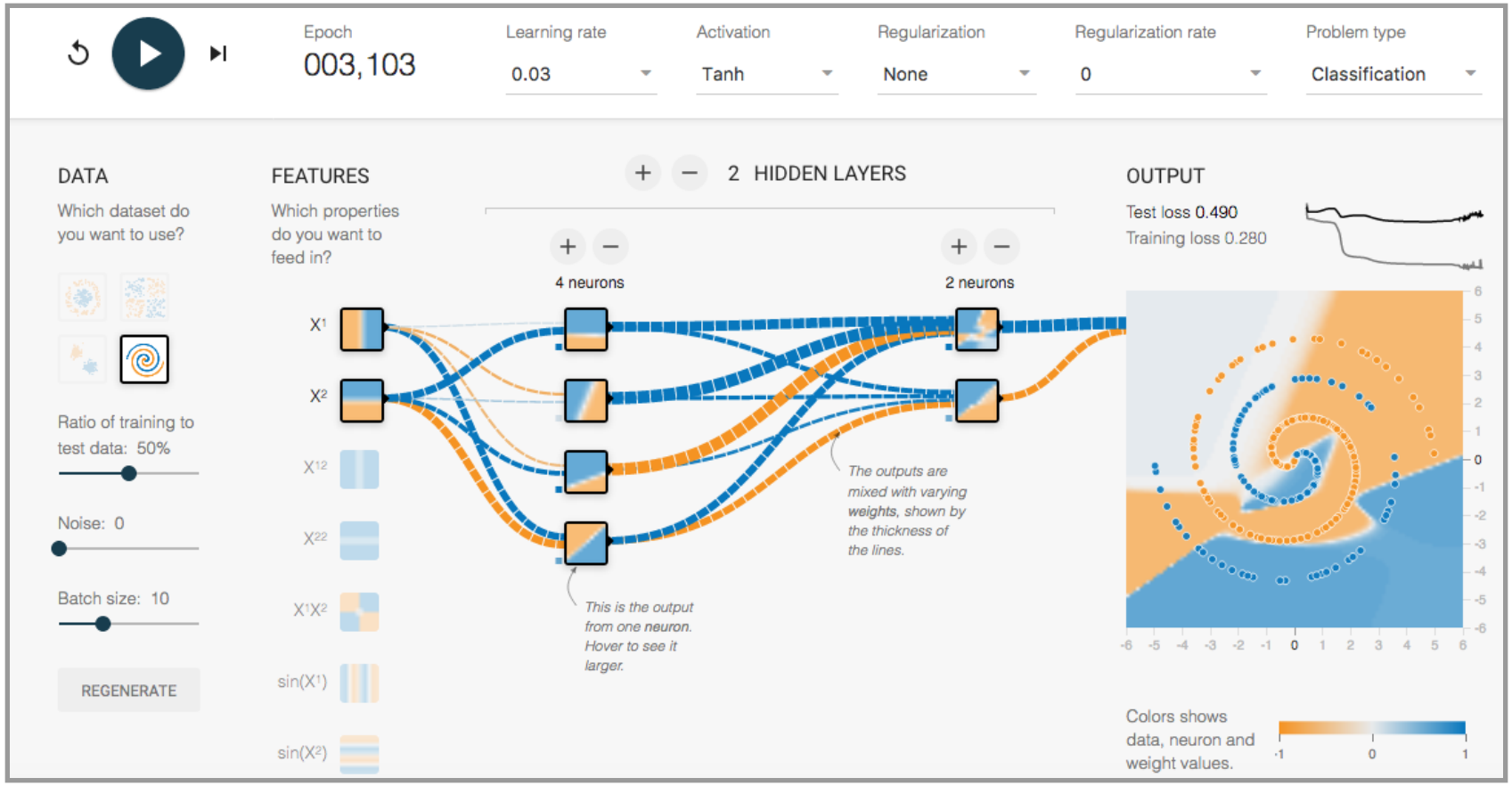
Universal Approximators

Theorem of Universal Approximation:
"Every bounded continuous function with bounded support can be approximated arbitrarily closely by a multi-layer perceptron by selecting enough but a finite number of hidden neurons with appropriate transfer function"

Cybenko., G. (1989) "Approximations by superpositions of sigmoidal functions", Mathematics of Control, Signals, and Systems, 2 (4), 303-314

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. Neural Networks, 2:359--366, 1989

ANN Playground



http://playground.tensorflow.org/

Workshop 1 – MLP for Regression

https://colab.research.google.com/drive/1yW5Ia5hxcaDRNtPGAG3dDb4UQ2PeXvz2?usp=sharing

MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

ihealth | Millennium Institute for Intelligent Healthcare Engineering

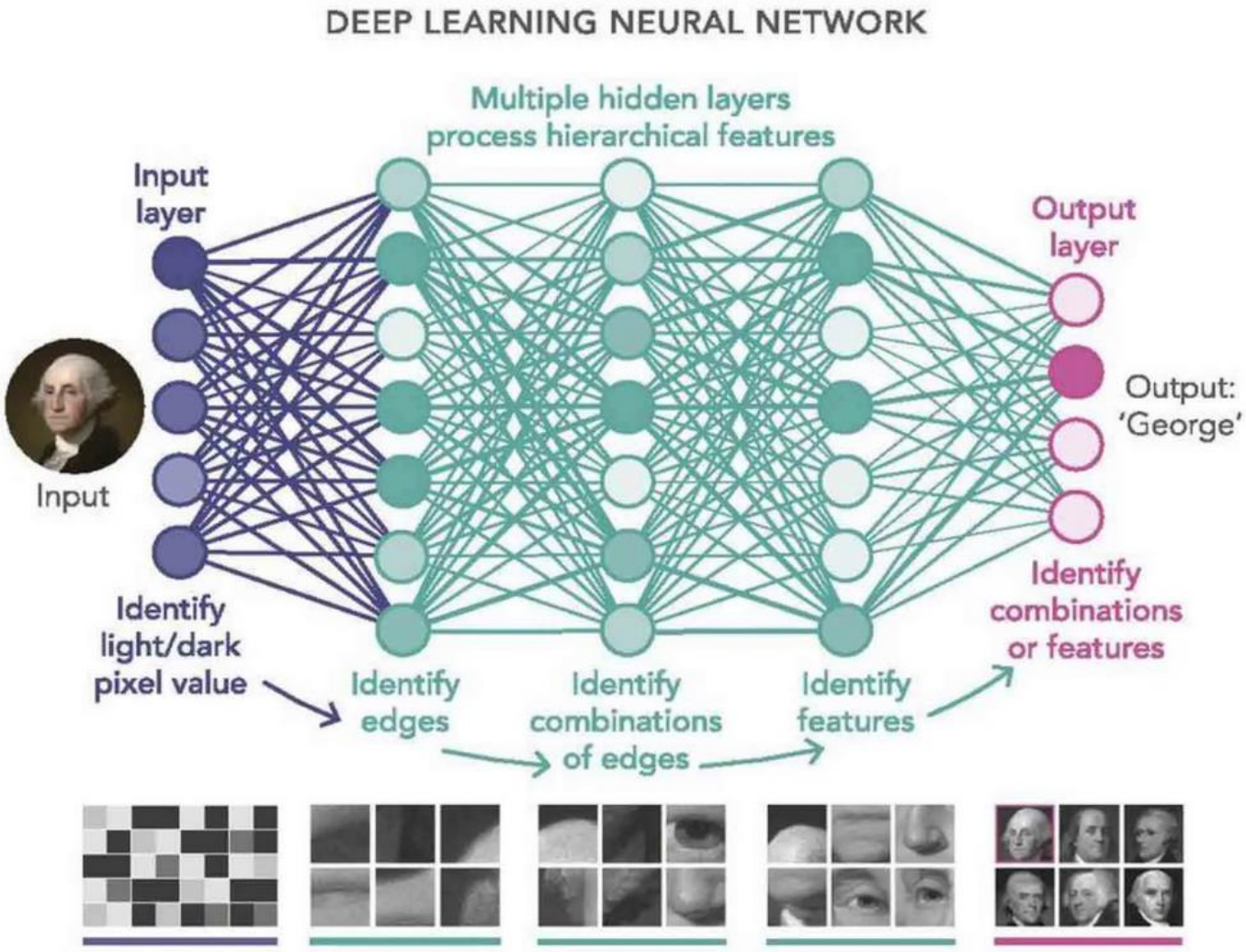
Deep Learning and Convolutional Neural Networks

CNN

Universidad de Valparaíso CHILE

¿What is Deep Learning?

- **Deep Learning (DL)** is a method based on deep neural networks that automatically learn complex structures and patterns from data through hierarchical and nonlinear representations.
- It is inspired by the architecture of the human brain.
- The model is composed of multiple intermediate layers.
- It enables the modeling of nonlinear relationships in large volumes of data.



M. Mitchell Warldrop, Artificial Neural Network; from https://www.pnas.org

Deep-Learning

Yann LeCun

Yoshua Bengio

Geoffrey Hinton

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 521 > Issue 7553 > Insights > Reviews > Article

ARTICLE PREVIEW

view full access options

NATURE | INSIGHT | REVIEW

Deep learning

Yann LeCun, Yoshua Bengio & Geoffrey Hinton

Affiliations | Corresponding author

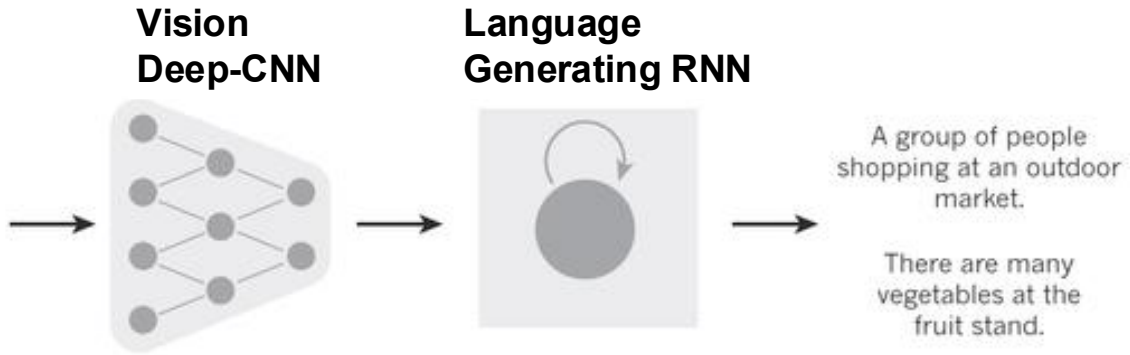
Nature **521**, 436–444 (28 May 2015) doi:10.1038/nature14539

Received 25 February 2015 | Accepted 01 May 2015 | Published online 27 May 2015

Citation | Reprints | Rights & permissions | Article metrics

Universidad de Valparaíso CHILE

High-level Translation from Images



A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background

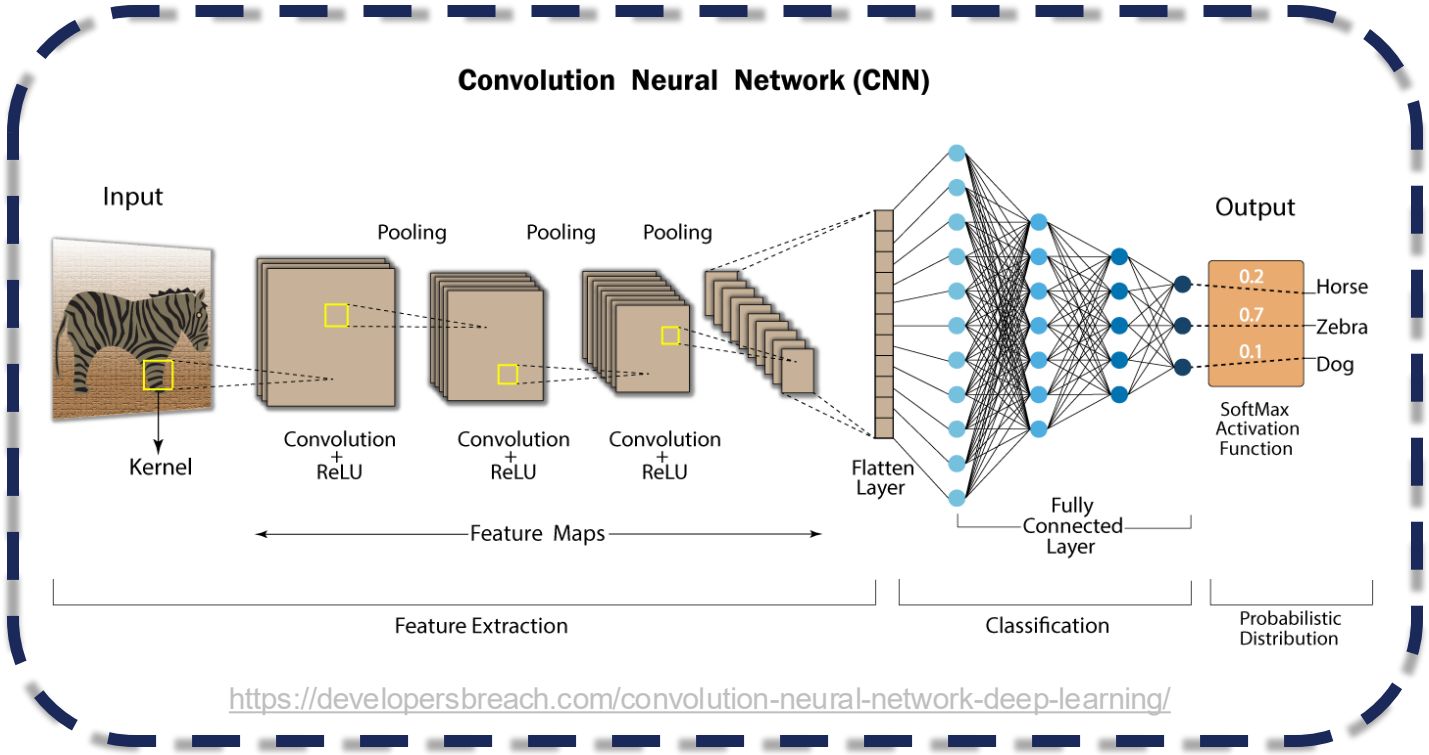
A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

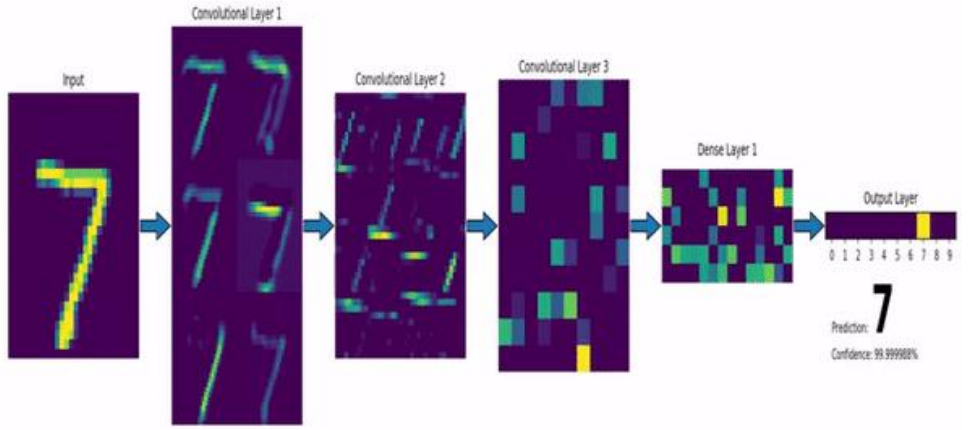
A giraffe standing in a forest with **trees** in the background.

LeCunn, Bengio & Hinton (2015). Deep learning. Nature 521, pp. 436–444. doi:10.1038/nature14539

Convolutional Neural Network



https://developersbreach.com/convolution-neural-network-deep-learning/

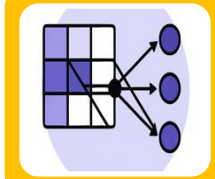


https://www.louisbouchard.ai/densenet-explained/

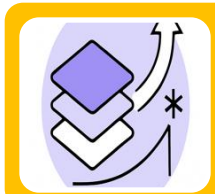
Convolutional Neural Network

A **Convolutional Neural Network (CNN)** is a type of neural network specialized in processing data with spatial structure. Its design enables the detection of visual patterns directly from the pixels, without the need to manually define features.

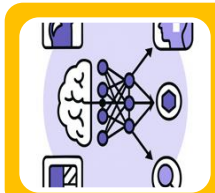
Each neuron connects only to a local region of the input (receptive field), capturing patterns such as edges or textures.



Layers are stacked hierarchically, learning from simple details to complex representations.

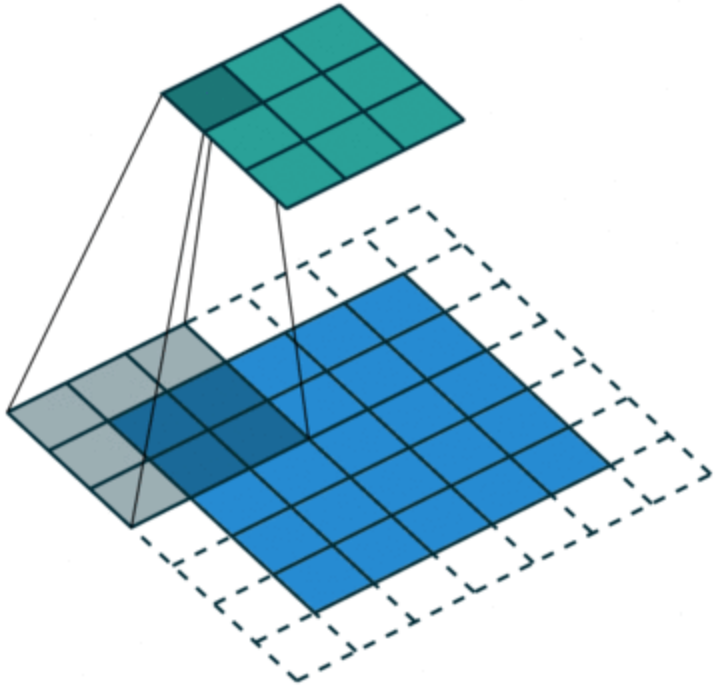


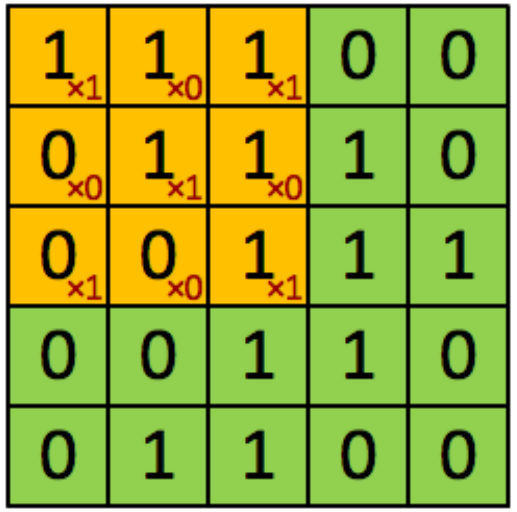
It uses **shared weights** and **filters** that slide across the image, drastically reducing the number of parameters.



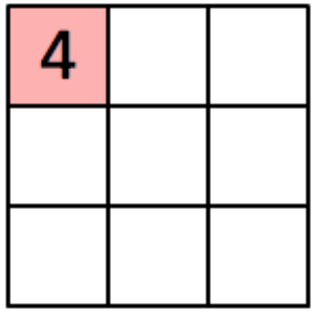
It is highly effective in tasks such as **classification, segmentation, and object detection**.

Convolutional Operations





Image

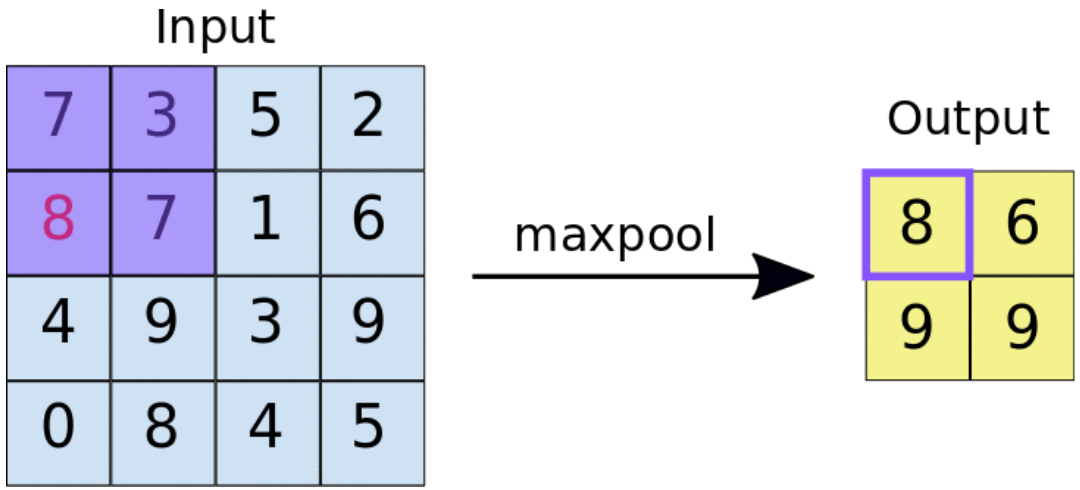


Convolved Feature

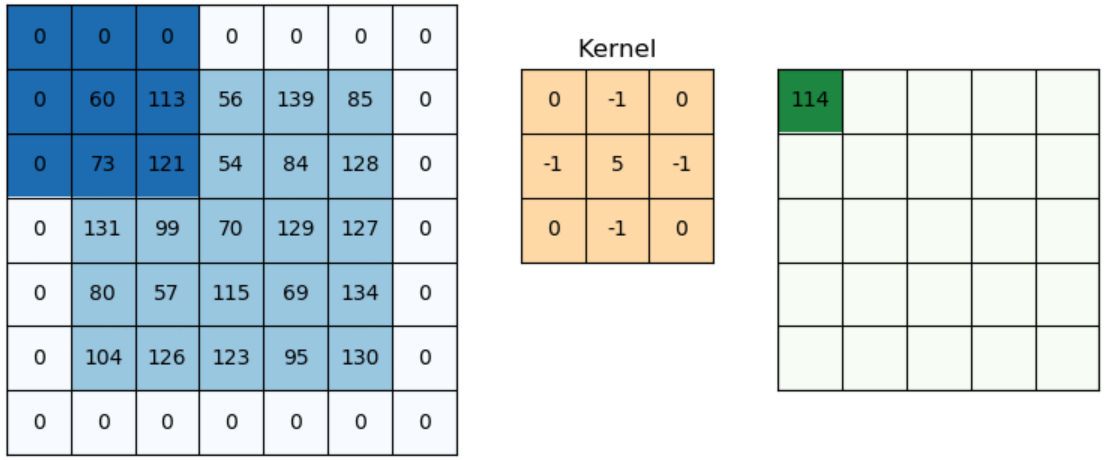
Kernel: A small filter that scans the image and extracts local features such as edges, textures, or patterns.

Convolution: An operation that multiplies and sums the kernel values with local regions of the image to generate a feature map.

Convolutional Operations



Max-Pooling: A dimensionality reduction technique that keeps only the maximum value from local regions, preserving the most important features.



Padding: Adding borders (usually zeros) around the image to preserve the output size or prevent loss of information at the edges.

Automatic Feature Extraction with Convolutional Layers

Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0	0	0	0	0	0	0
0	0	0	1	0	2	0
0	1	0	2	0	1	0
0	1	0	2	2	0	0
0	2	0	0	2	0	0
0	2	1	2	2	0	0
0	0	0	0	0	0	0

x[:,:,1]

0	0	0	0	0	0	0
0	2	1	2	1	1	0
0	2	1	2	0	1	0
0	0	2	1	0	1	0
0	1	2	2	2	2	0
0	0	1	2	0	1	0
0	0	0	0	0	0	0

x[:,:,2]

0	0	0	0	0	0	0
0	2	1	1	2	0	0
0	1	0	0	1	0	0
0	0	1	0	0	0	0
0	1	0	2	1	0	0
0	2	2	1	1	1	0
0	0	0	0	0	0	0

Filter W0 (3x3x3)

w0[:,:,0]

-1	0	1
0	0	1
1	-1	1

w0[:,:,1]

-1	0	1
1	-1	1
0	1	0

w0[:,:,2]

-1	1	1
1	1	0
0	-1	0

Bias b0 (1x1x1)

b0[:,:,0]

1

Filter W1 (3x3x3)

w1[:,:,0]

0	1	-1
0	-1	0
0	-1	1

w1[:,:,1]

-1	0	0
1	-1	0
1	-1	0

w1[:,:,2]

-1	1	-1
0	-1	-1
1	0	0

Bias b1 (1x1x1)

b1[:,:,0]

0

Output Volume (3x3x2)

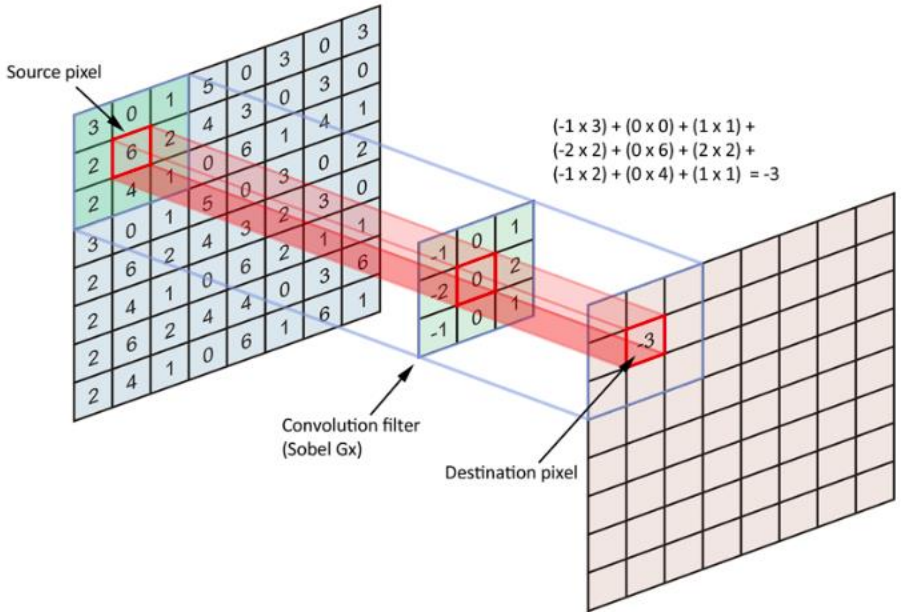
o[:,:,0]

2	3	3
3	7	3
8	10	-3

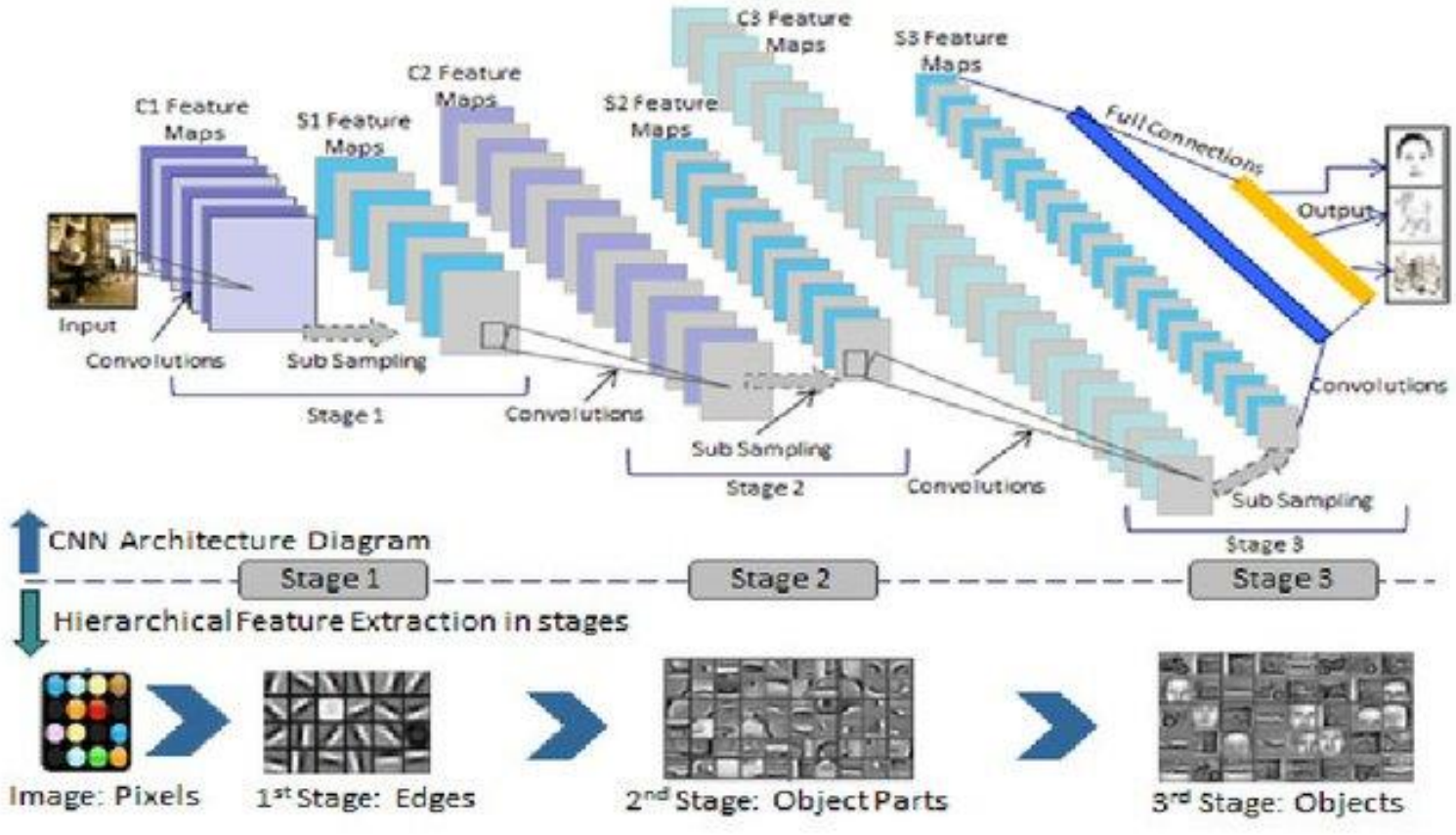
o[:,:,1]

-8	-8	-3
-3	1	0
-3	-8	-5

toggle movement



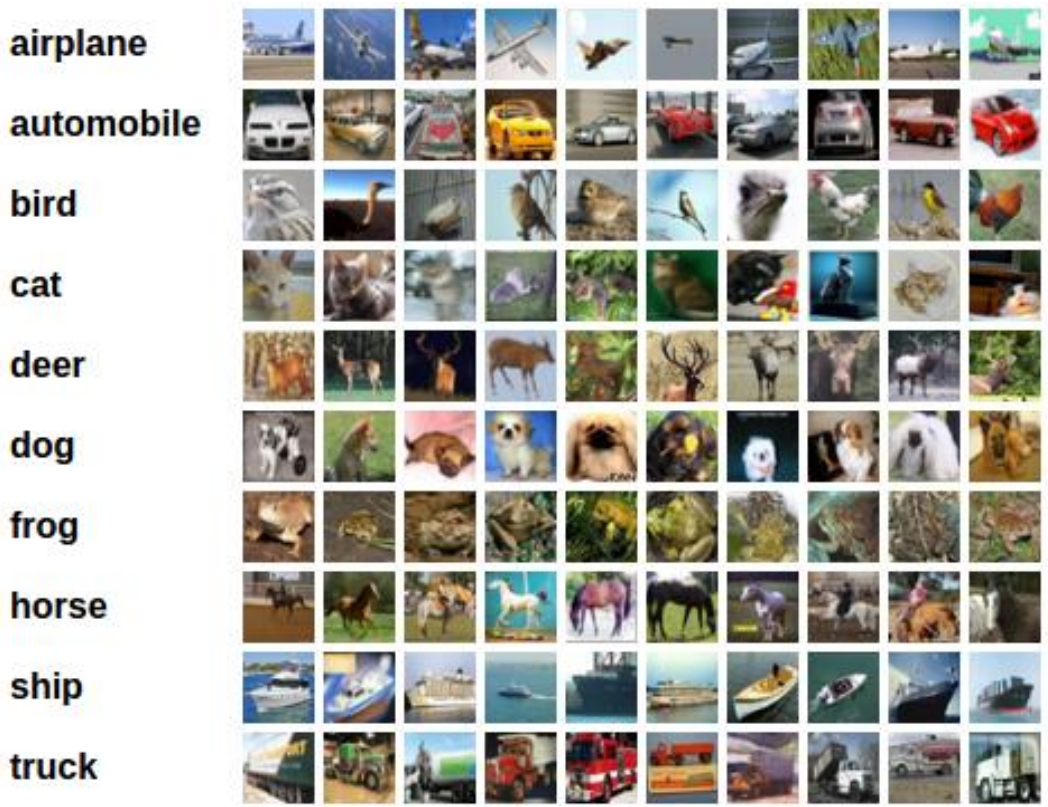
Learning Hierarchy of Visual Features in CNN Architecture

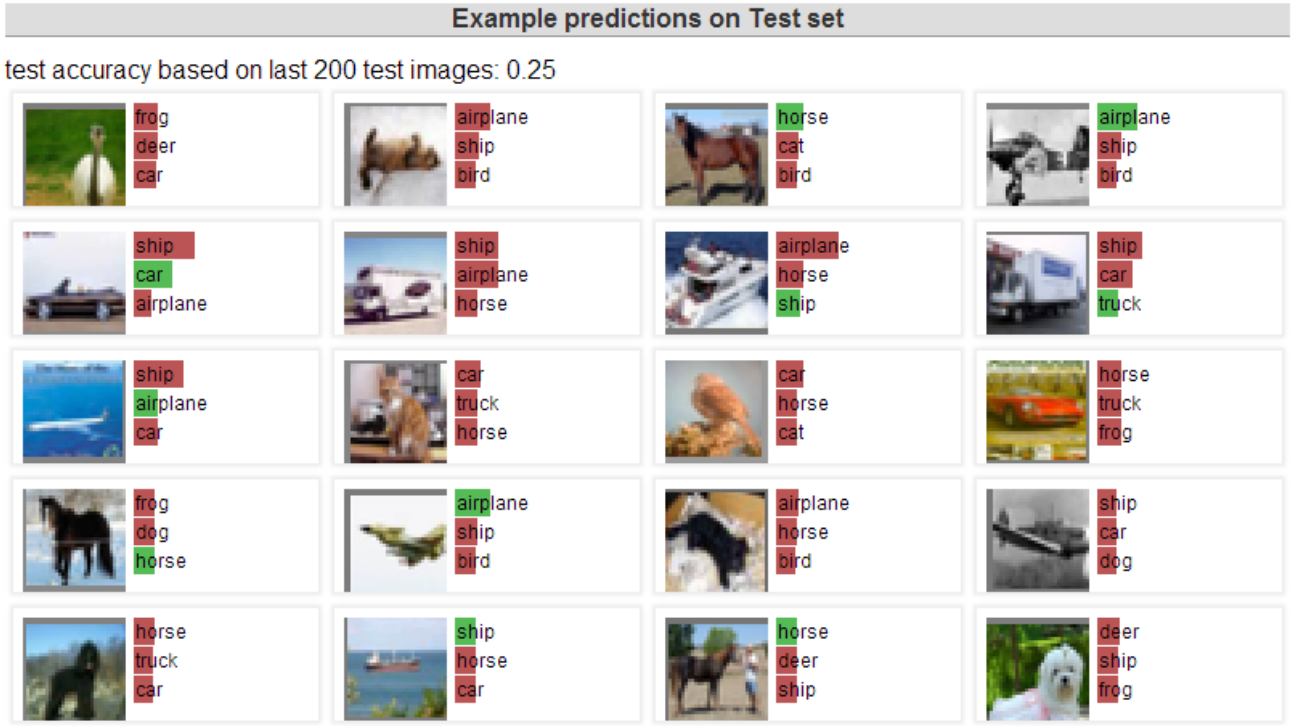


Katole et al. Hierarchical Deep Learning Architecture For 10K Objects Classification. Doi: 10.5121/csit.2015.51408

ConvNetJS CIFAR-10 demo

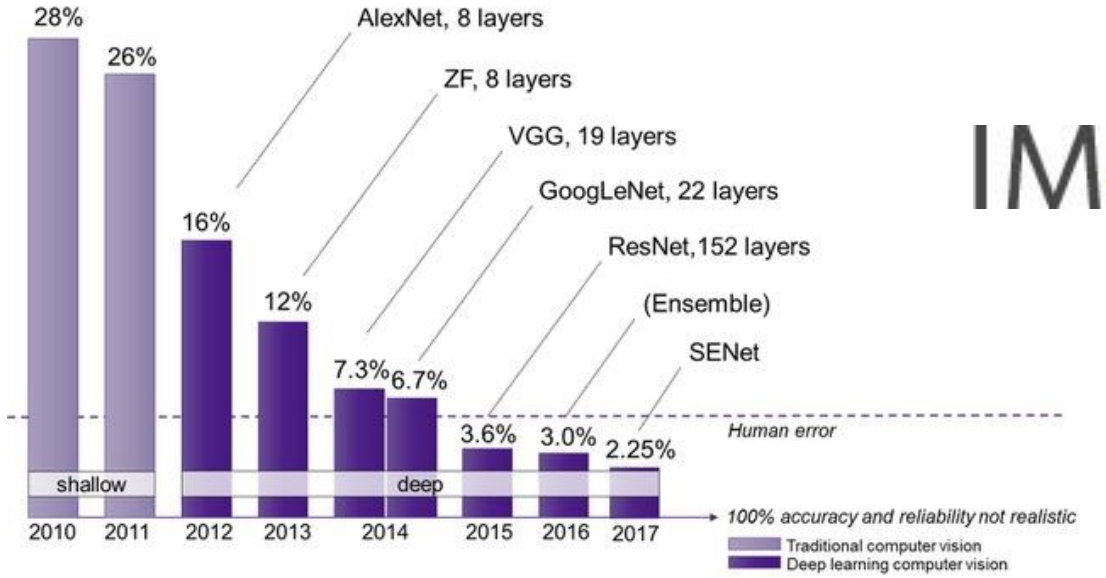
ConvNetJS
Deep Learning in your browser



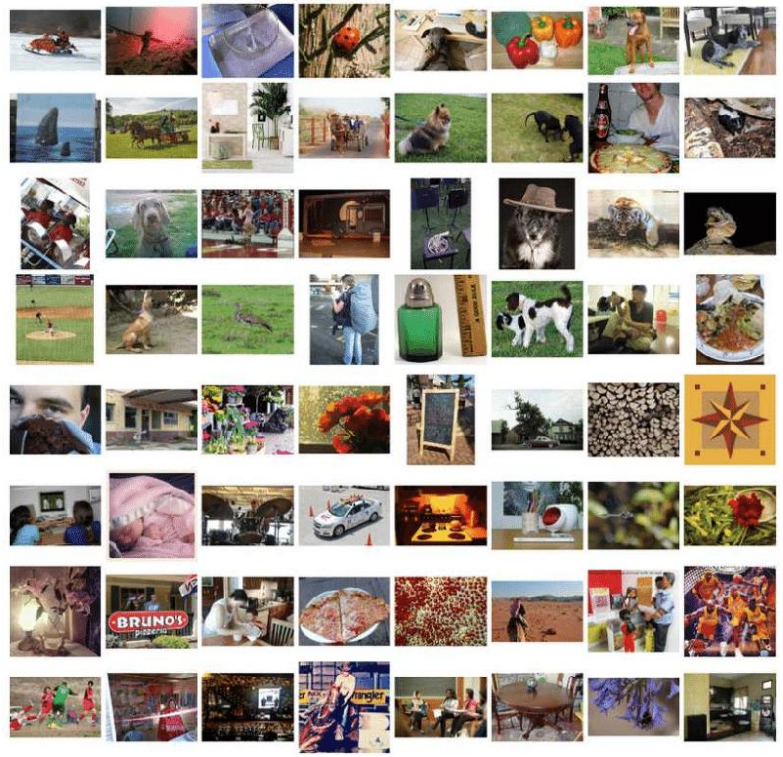


http://cs.stanford.edu/people/karpathy/convnetjs/demo/cifar10.html

Large Scale Visual Recognition Challenge (ILSVRC)



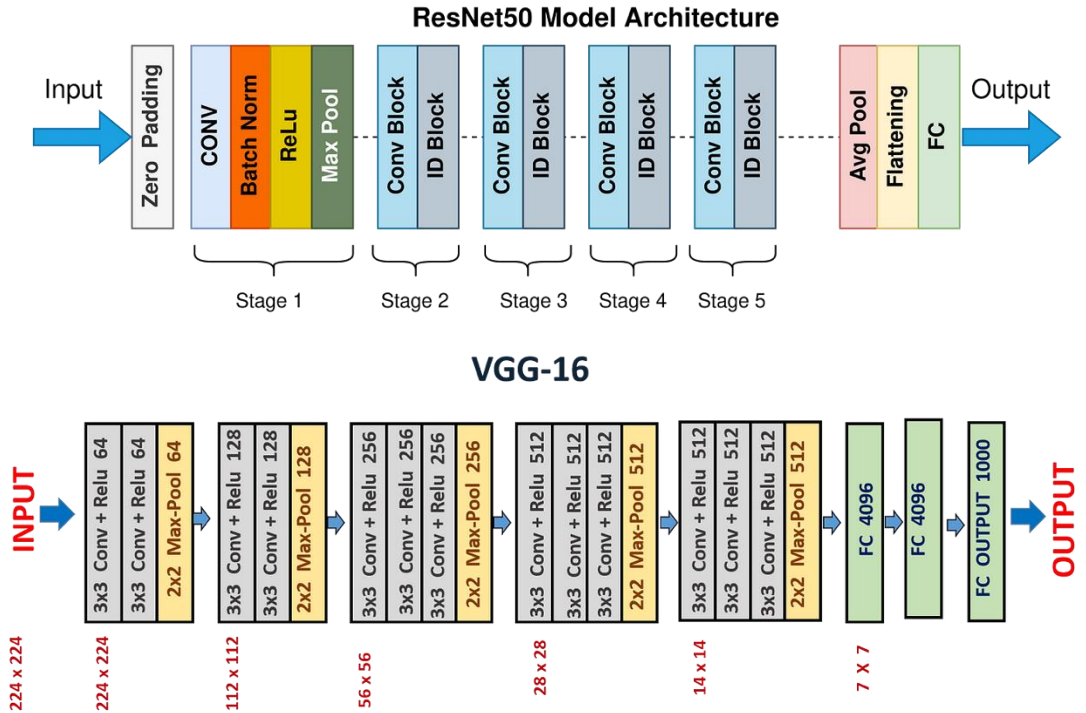
IMAGENET



http://www.image-net.org/challenges/LSVRC/

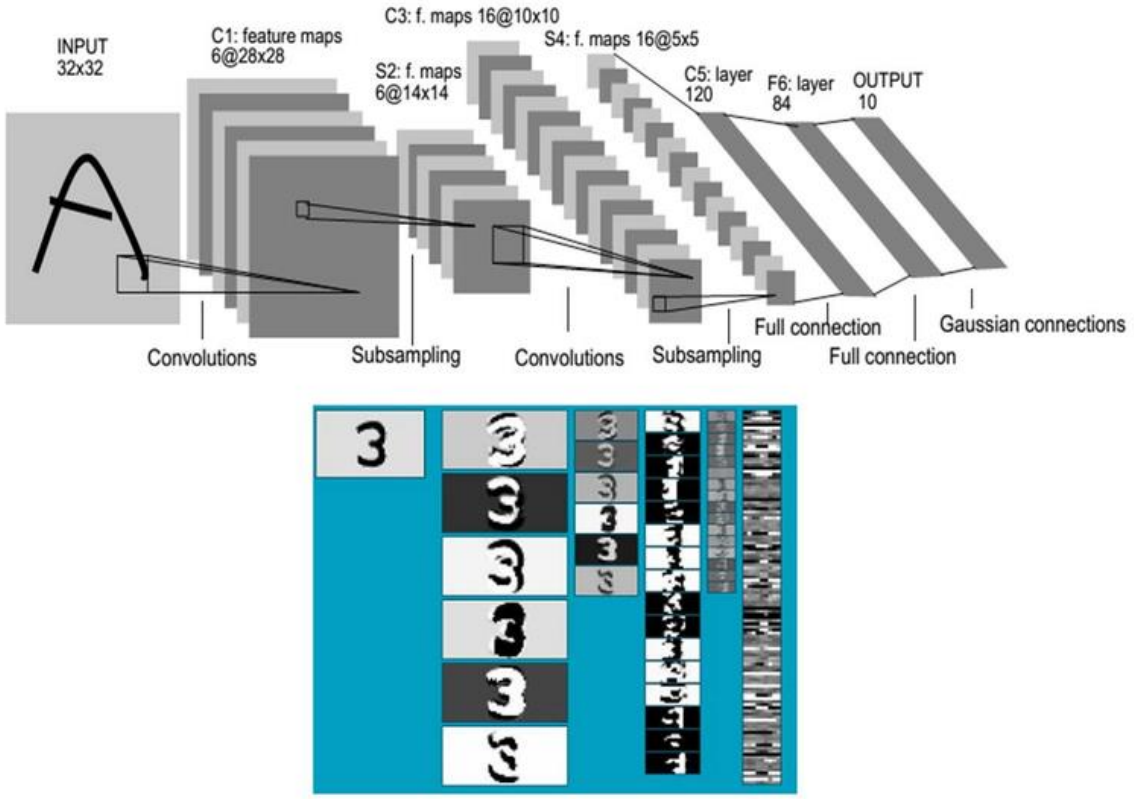
https://gigaom.com/2014/08/22/with-enlitic-a-veteran-data-scientist-plans-to-fight-disease-using-deep-learning/

Architectures of CNN

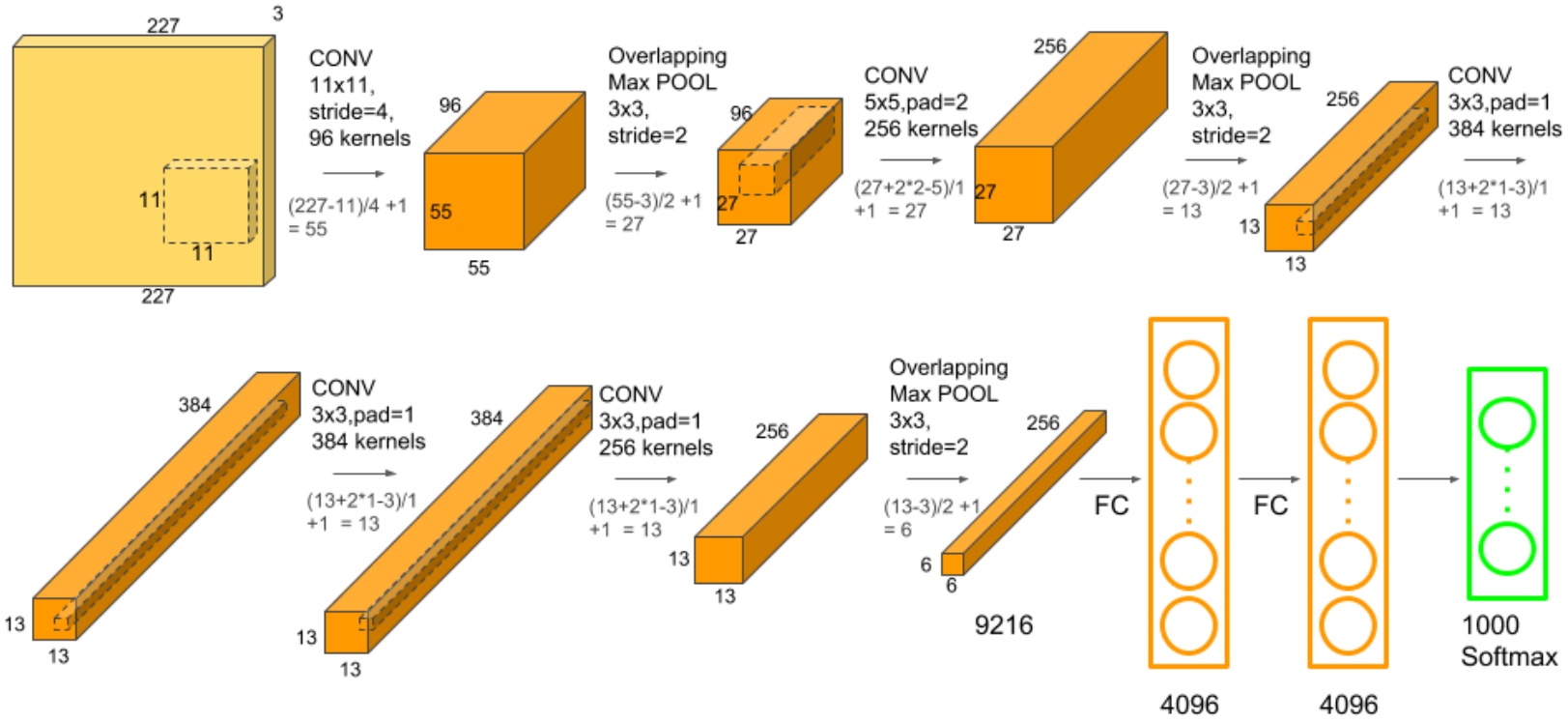


Estas arquitecturas CNN (VGG, ResNet, U-Net) aprenden representaciones visuales jerárquicas y se adaptan a tareas como clasificación y segmentación pixel a pixel.

LeNet

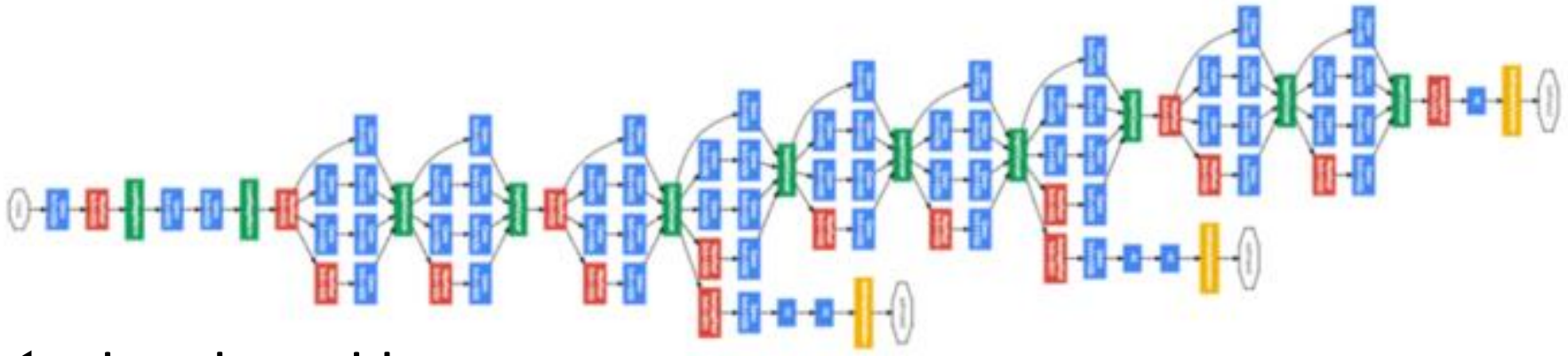

INPUT
32x32
C1: feature maps
6@28x28
S2: f. maps
6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
C5: layer
120
F6: layer
84
OUTPUT
10
Convolutions
Subsampling
Convolutions
Subsampling
Full connection
Full connection
Gaussian connections

AlexNet

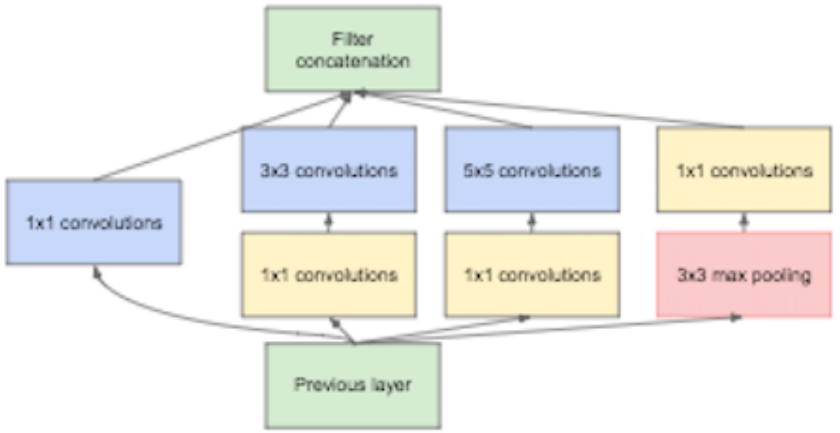


https://medium.com/@smallfishbigsea/a-walk-through-of-alexnet-6cbd137a5637

GoogLeNet



- **GoogLeNet:** developed by C. Szegedy et al. from Google. (2014).
- Introduced the **Inception module**, which reduced the number of parameters in the network (**sparsity**).

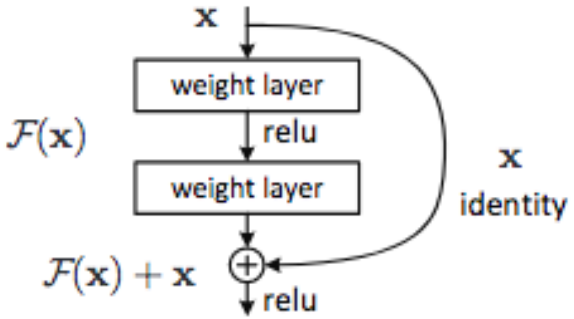


Módulo de Inception con reducción de la dimensionalidad

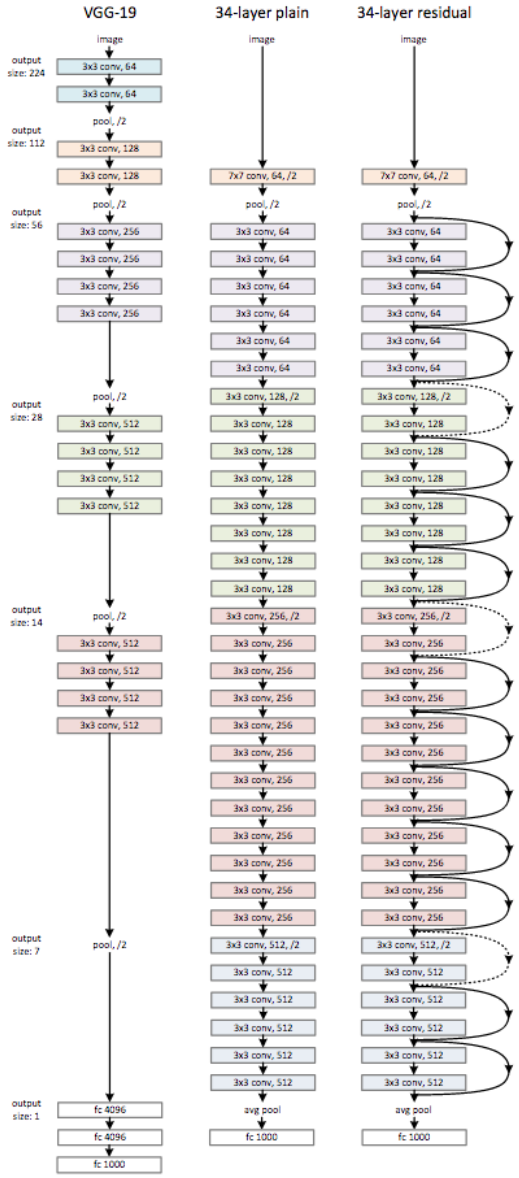
Residual Learning

- **Residual learning (ResNet)** was proposed by *K. He et al.* (2015).

- It is used to enable learning at greater depth.

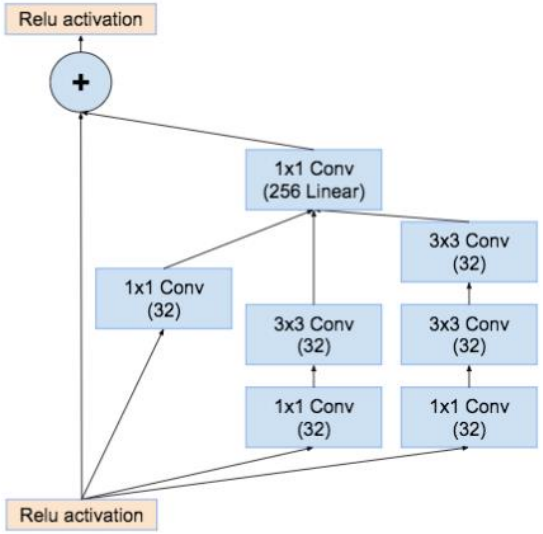


Bloque de aprendizaje residual

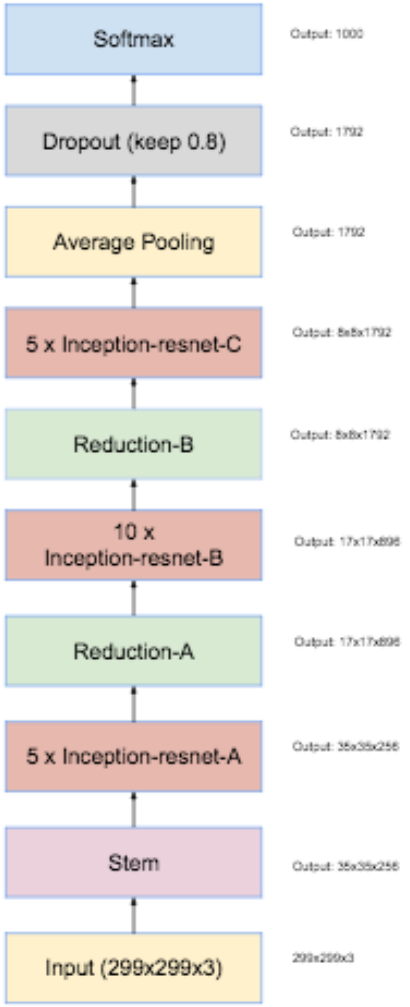


Inception-Resnet

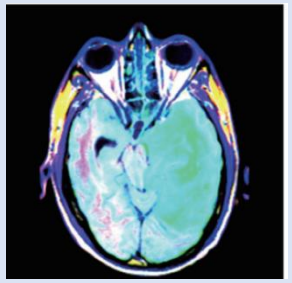
- Proposed by C. Szegedy 2016
- Combines the Inception module (Sparsity) with the ResNet module (Deep)



Módulo de Inception-ResNet-A

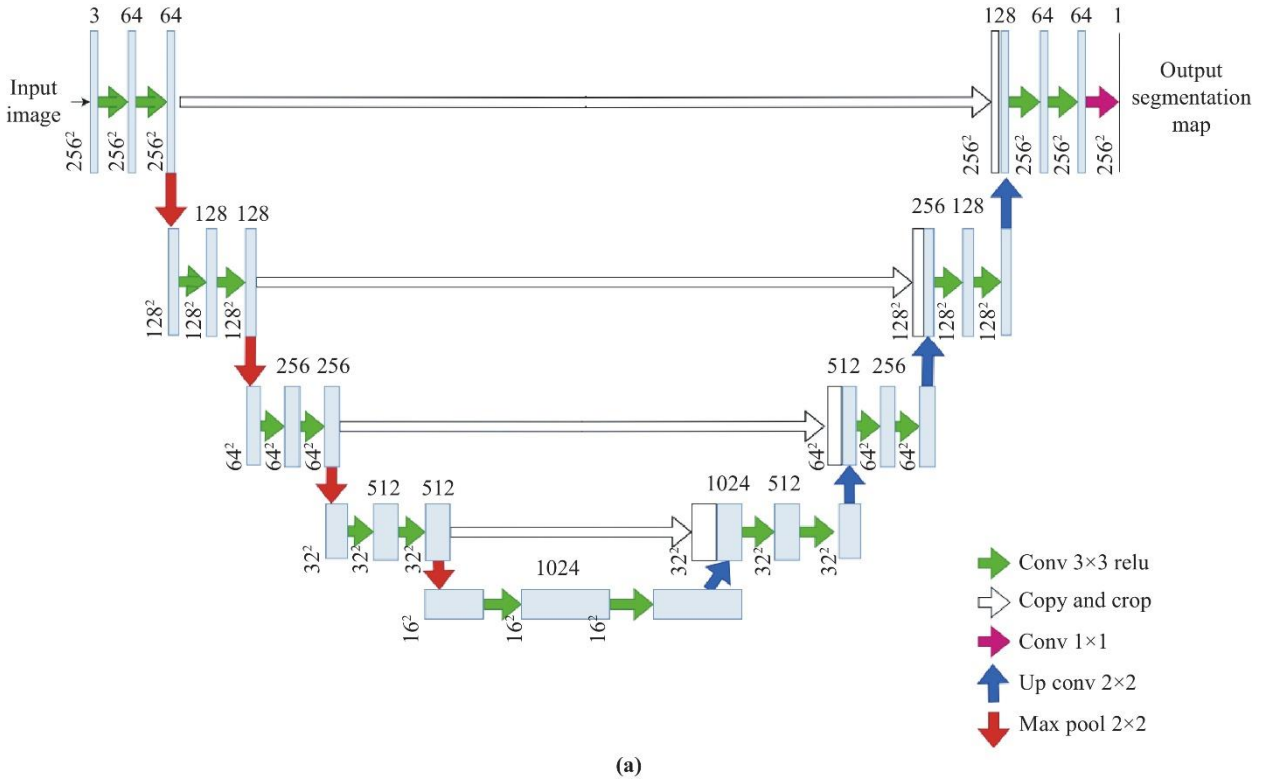


U-Net: Convolutional Networks for Biomedical Image Segmentation



Input Image

Ground Truth



Segmentation result

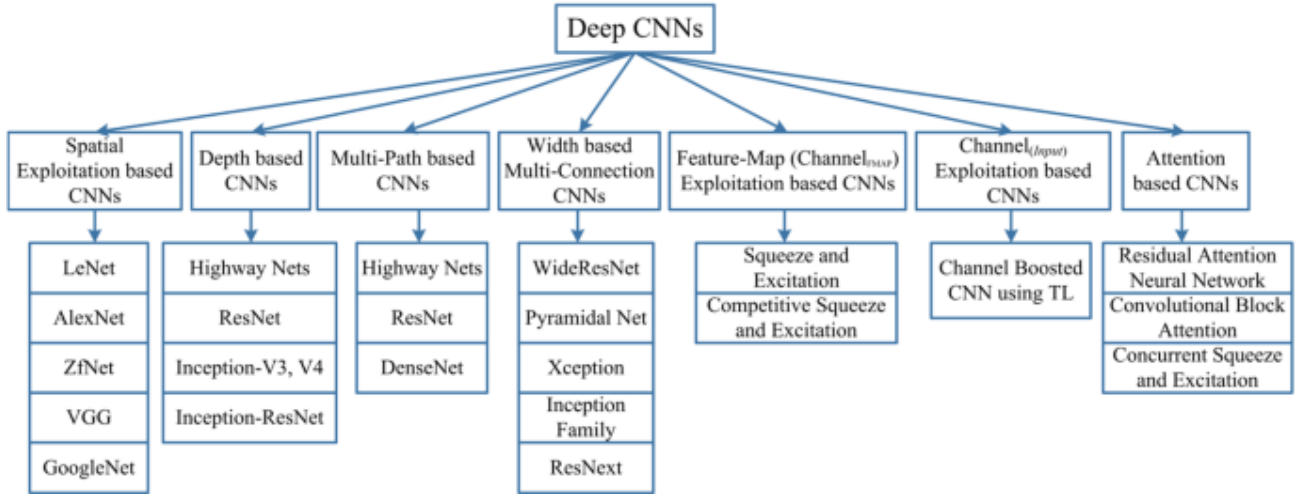
https://ojs.sciltp.com/journals/ijndi/article/view/411

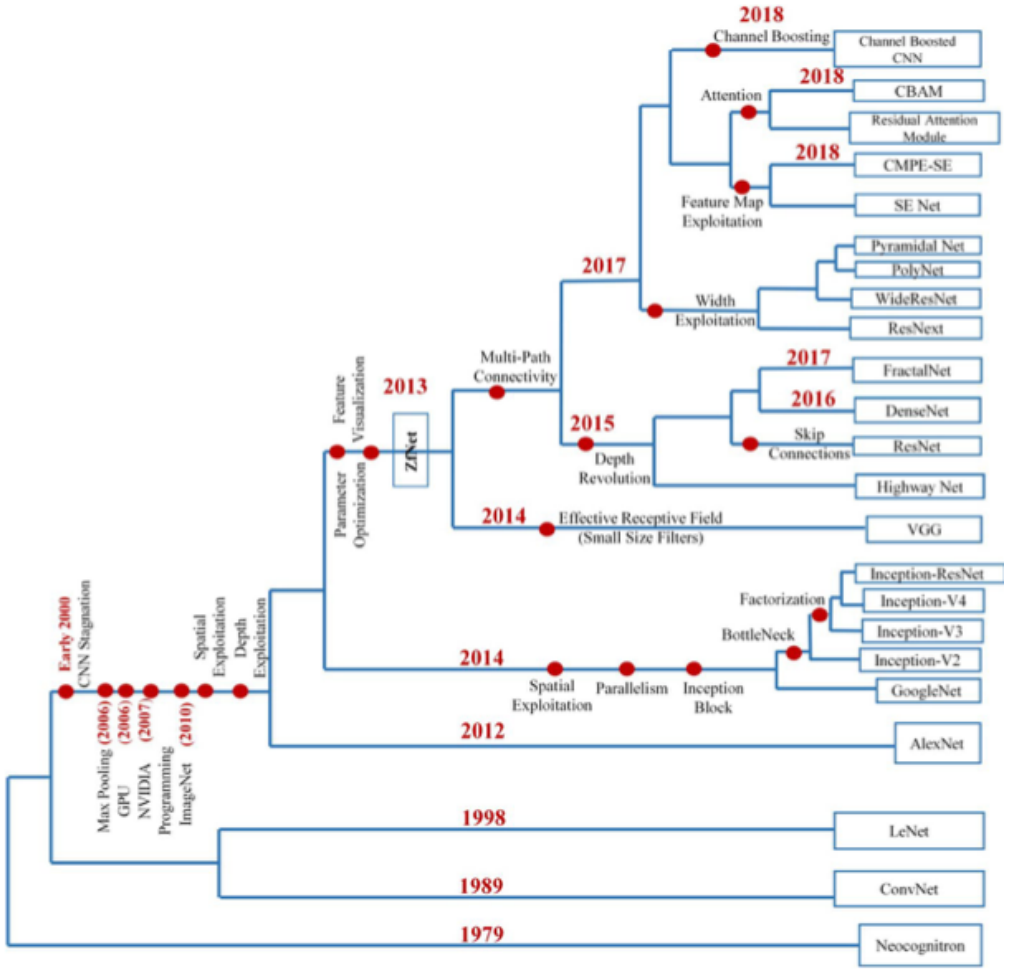
Architectures of Deep Convolutional Neural Networks

Artificial Intelligence Review (2020) 53:5455–5516
https://doi.org/10.1007/s10462-020-09825-6

A survey of the recent architectures of deep convolutional neural networks

Asifullah Khan[1,2] · Anabia Sohail[1,2] · Umme Zahoora[1] · Aqsa Saeed Qureshi[1]





Benchmark Analysis of Deep Learning

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 1, 2018, accepted October 17, 2018, date of publication October 24, 2018, date of current version November 19, 2018.

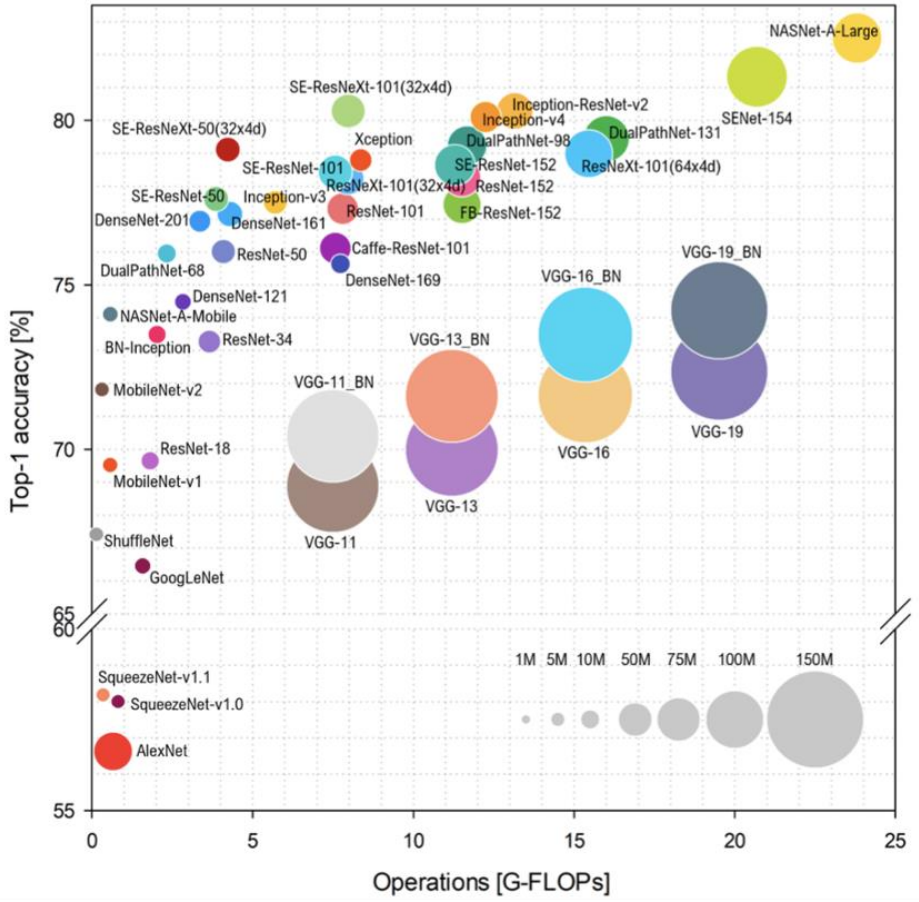
Digital Object Identifier 10.1109/ACCESS.2018.2877890

Benchmark Analysis of Representative Deep Neural Network Architectures

SIMONE BIANCO[1], REMI CADENE[2], LUIGI CELONA[1], AND PAOLO NAPOLETANO[1]
[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy
[2]LIP6, CNRS, Sorbonne Université, 75005 Paris, France
Corresponding author: Luigi Celona (luigi.celona@disco.unimib.it)

This paper presents an in-depth analysis of the majority of the deep neural networks (DNNs) proposed in the state of the art for image recognition.

For each DNN, multiple performance indices are observed, such as recognition accuracy, model complexity, computational complexity, memory usage, and inference time.



Workshop 2 – CNN applied for MNIST

https://colab.research.google.com/drive/18Hr-Bz6W1FcgCe67vXwdKviPn8W8IbjI?usp=sharing

MNIST Web Demo

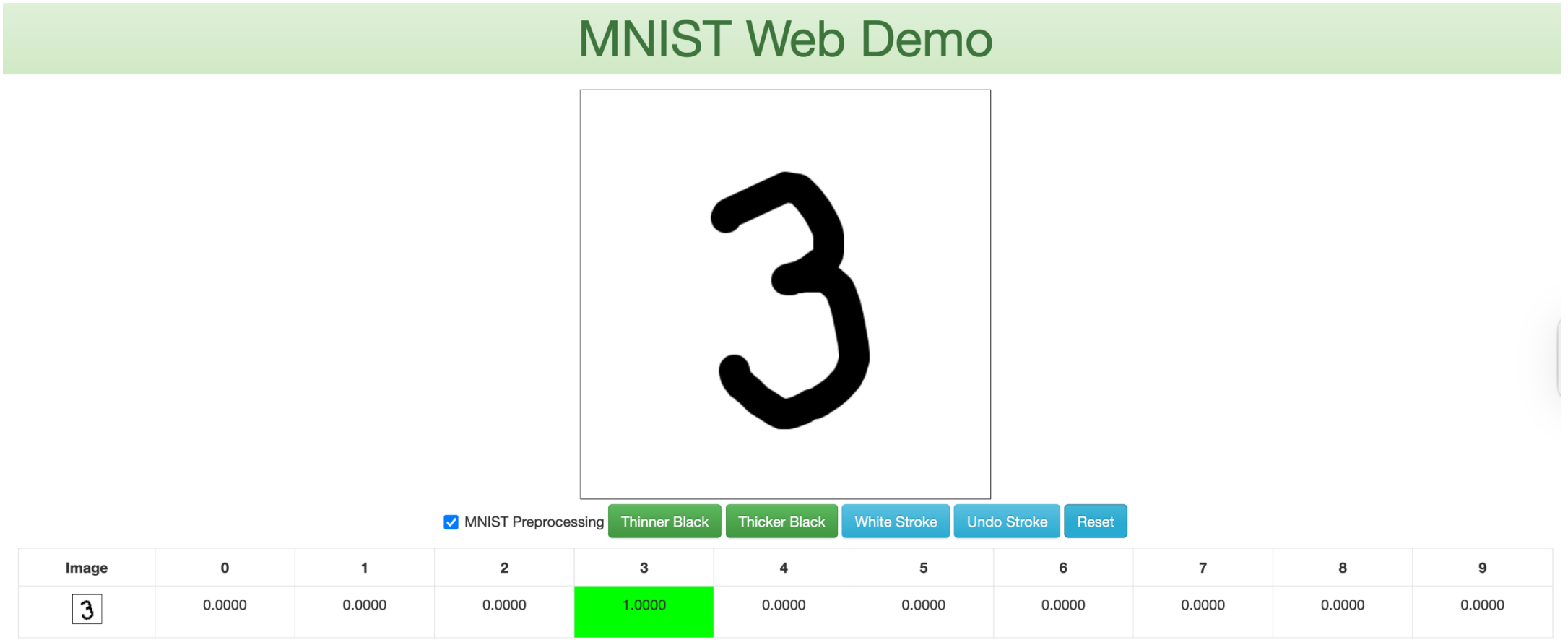


Image	0	1	2	3	4	5	6	7	8	9
3	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

https://ufal.mff.cuni.cz/~courses/npfl129/2425/demos/mnist_web.html

Workshop 3 – CNN Applied to the Sign Language MNIST Dataset

https://www.kaggle.com/code/rodsalasf/simposio-estadistica-taller-3-cnn

https://www.kaggle.com/code/madz2000/cnn-using-keras-100-accuracy

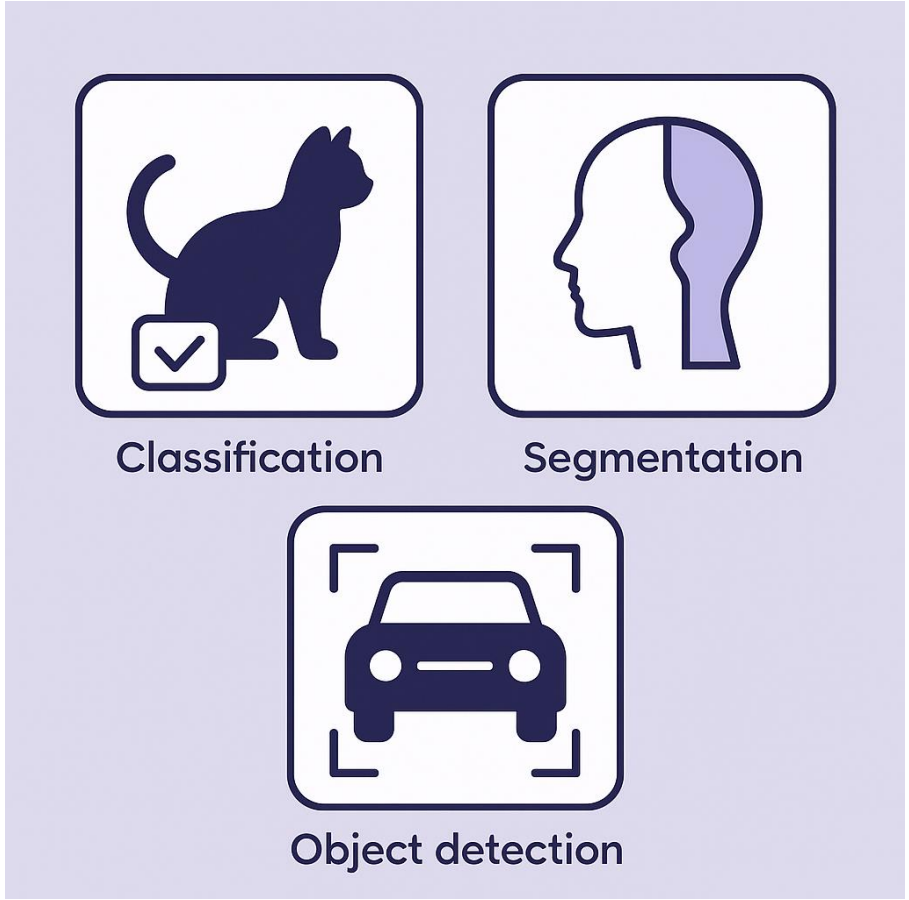
MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

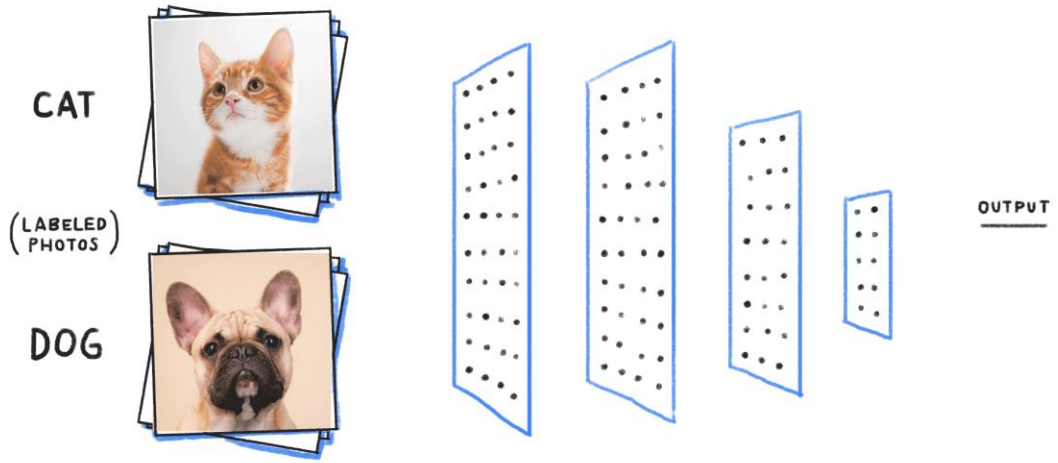
ihealth | Millennium Institute for Intelligent Healthcare Engineering

Deep Learning Applied for Image Processing

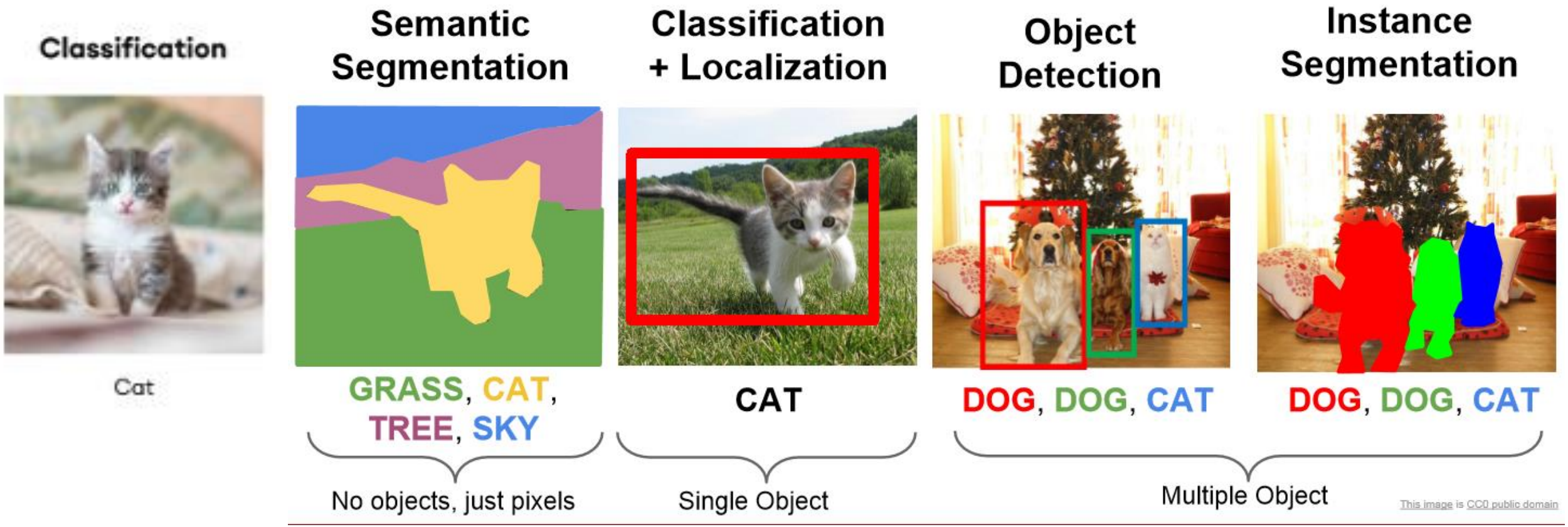
Universidad de Valparaíso CHILE

Deep Learning applied for Image Processing





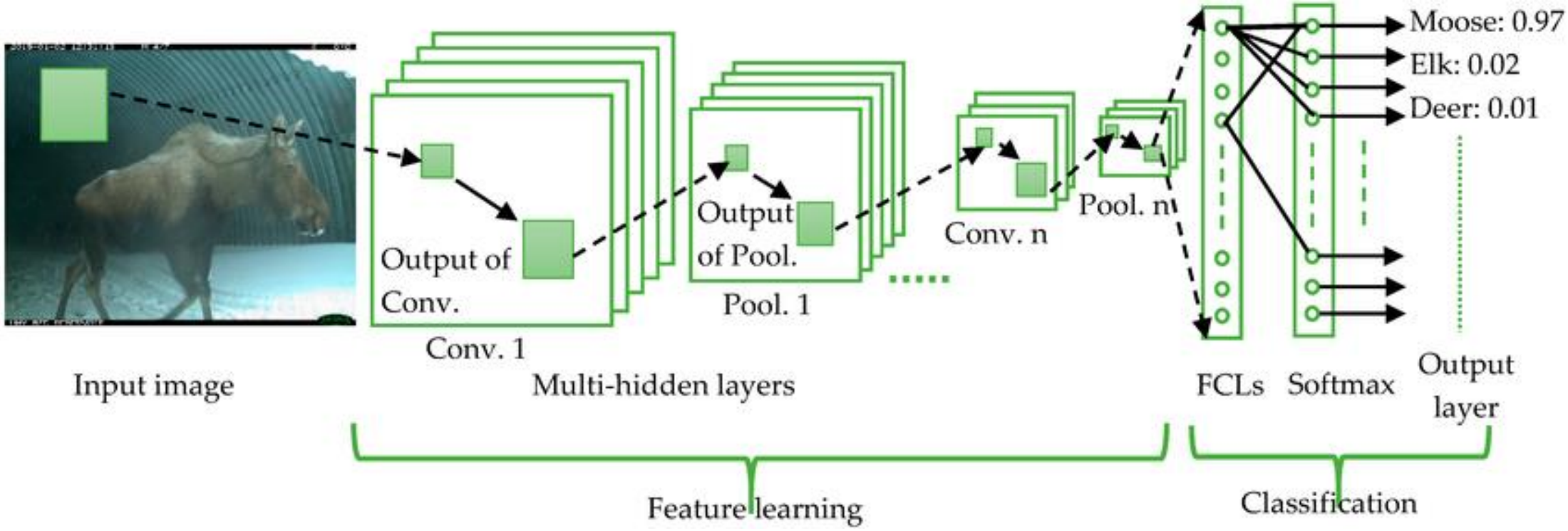
Deep Learning Tasks in Computer Vision



Comparison of semantic segmentation, classification and localization, object detection and instance segmentation (Li, Johnsonand Yeung, 2017)

https://www.superannotate.com/blog/image-segmentation-for-machine-learning

Image Classification

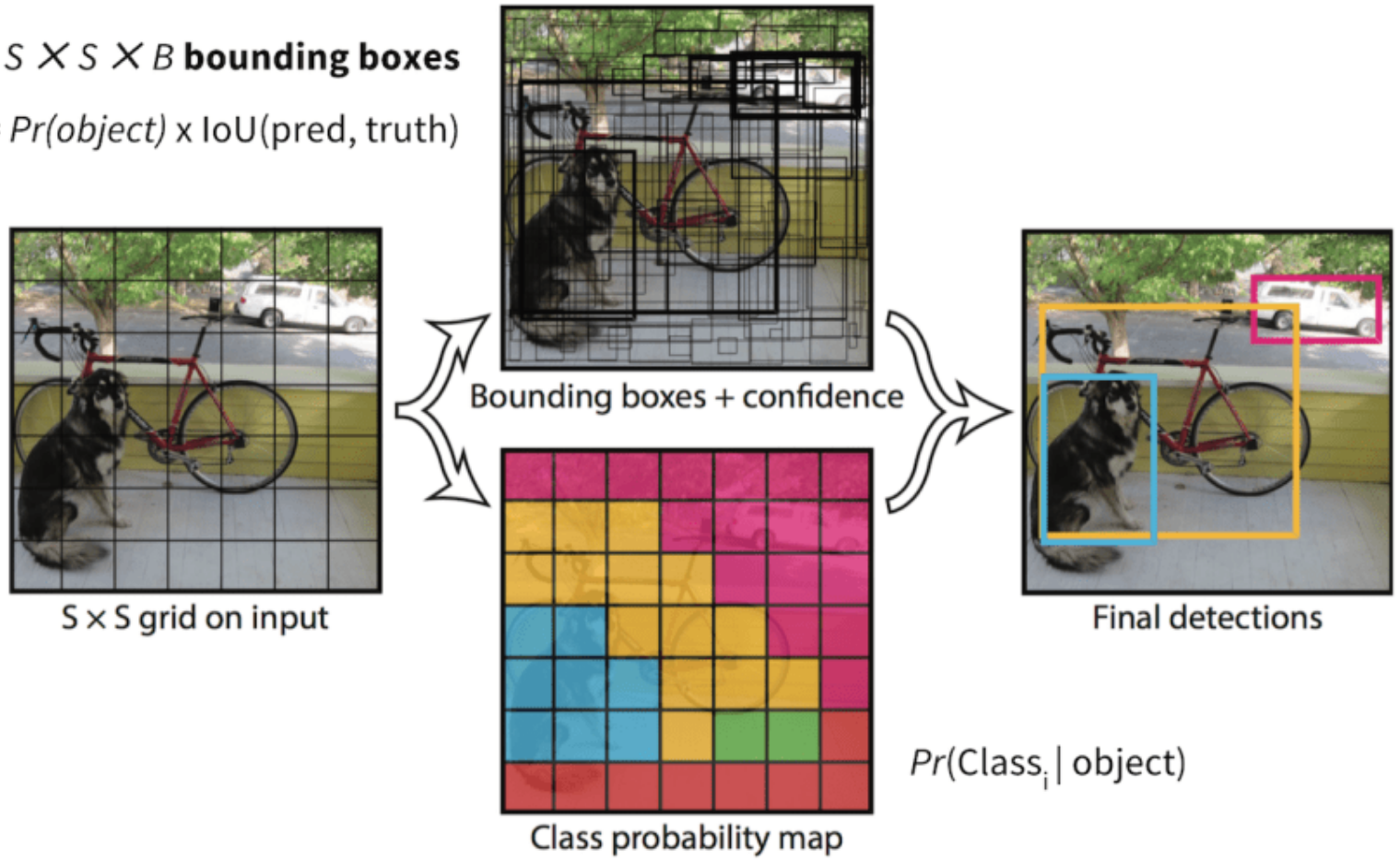


The network extracts **visual features** from the image through convolution and pooling layers. These representations are then processed in fully connected layers to assign probabilities to each class. The model learns to identify patterns that allow it to automatically classify the content of the image.

Object Detection

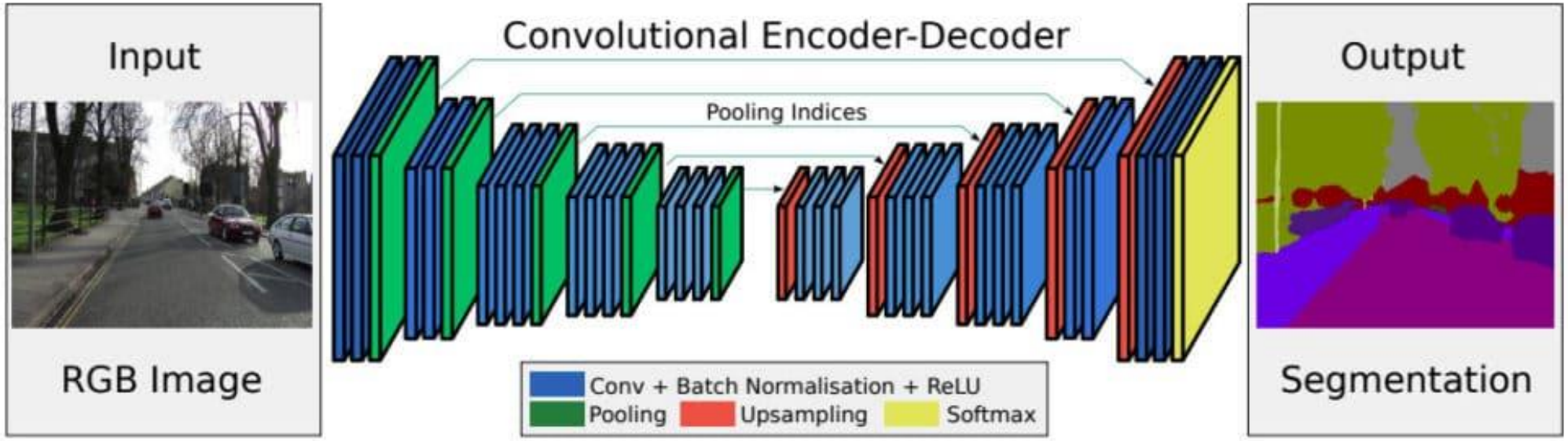
$S \times S \times B$ **bounding boxes**

confidence = *Pr(object)* x IoU(pred, truth)



Object detection combines localization and classification: the model divides the image into a grid, predicts bounding boxes, and assigns a class probability to each one. Using metrics such as **IoU** and **confidence scores**, multiple objects can be identified and located within a single image.

Semantic Segmentation



Semantic segmentation assigns a label to every pixel in the image, identifying which class each region belongs to (e.g., street, car, person).

This process is performed using **encoder–decoder networks**, where the encoder extracts features and the decoder reconstructs a class map with the same resolution as the original image.

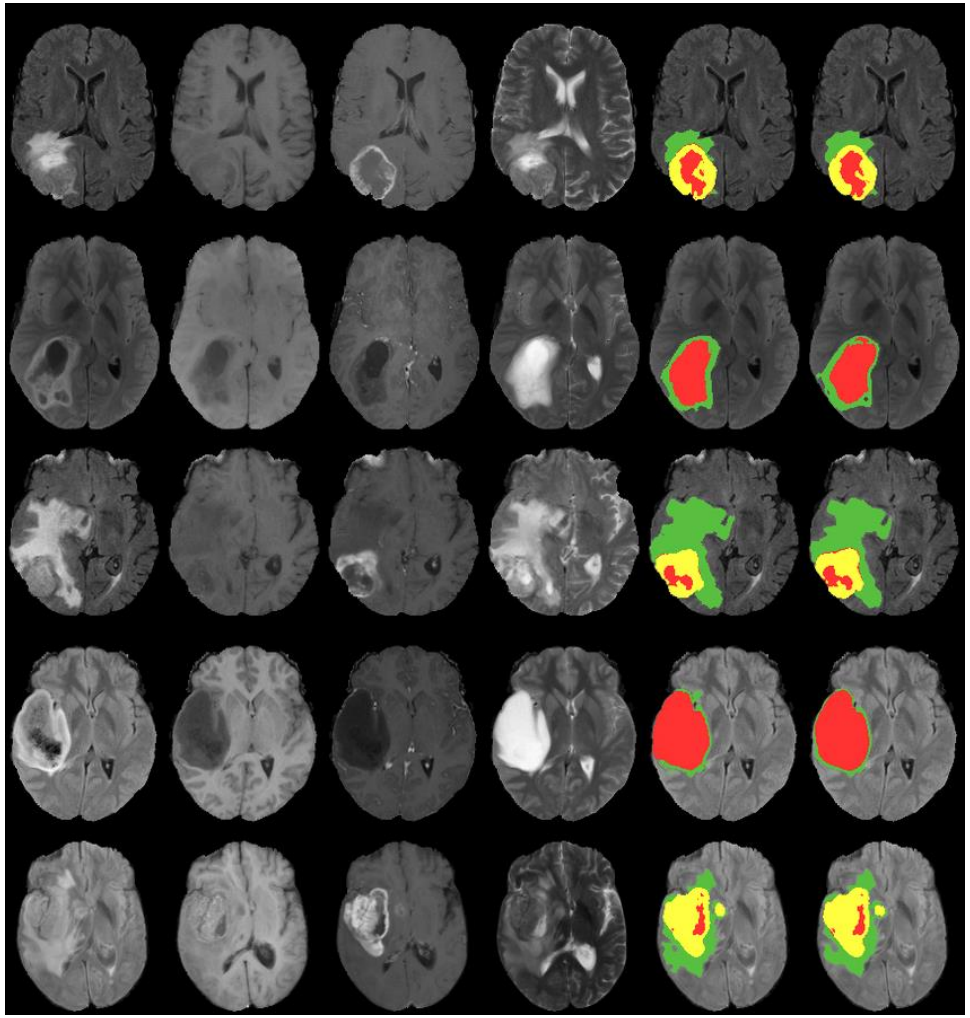
U-NET Applied for Semantic Segmentation

MSc. Student
Gabriel Guerra

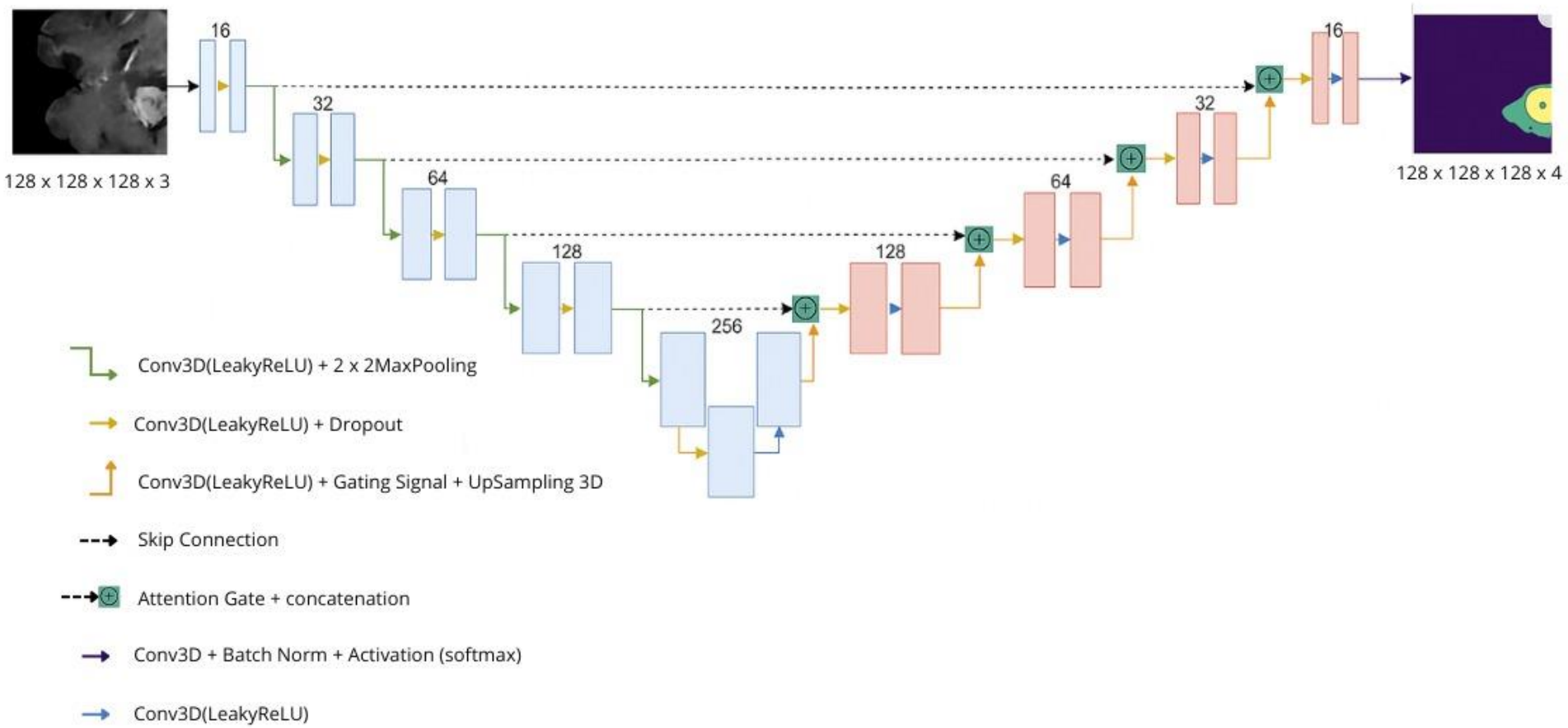
Material desarrollado con el apoyo del estudiante del Magister en Ciencias e Ingeniería para la Salud, Universidad de Valparaíso
Sr. Gabriel Guerra

Segmentación de tumores cerebrales

BraTS es un conjunto de datos público que contiene resonancias magnéticas cerebrales multimodales (T1, T1c, T2, FLAIR) de pacientes con gliomas. Incluye segmentaciones manuales de tres regiones tumorales: tumor completo, tumor realzado y tumor no realzado. Su objetivo es evaluar algoritmos de segmentación automática en contextos clínicos realistas.

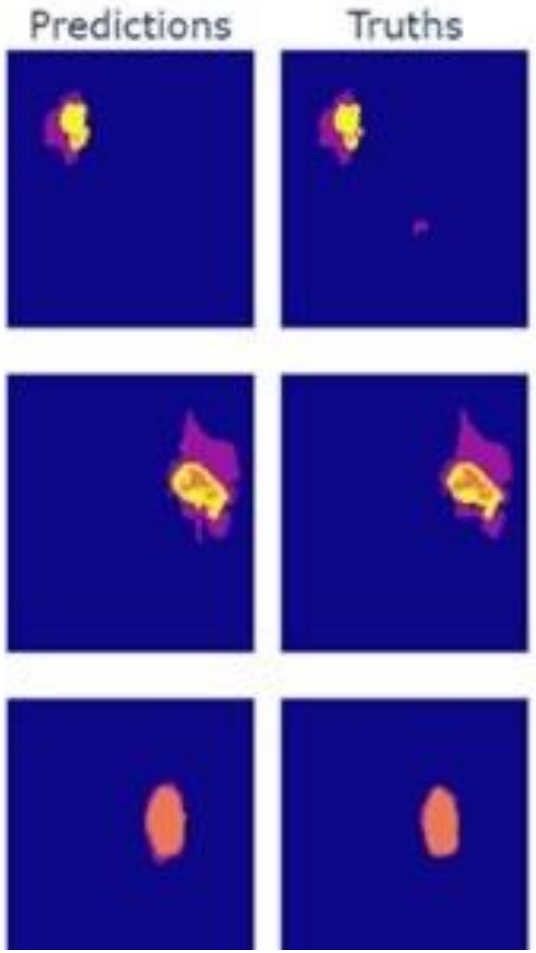


Attention U-net

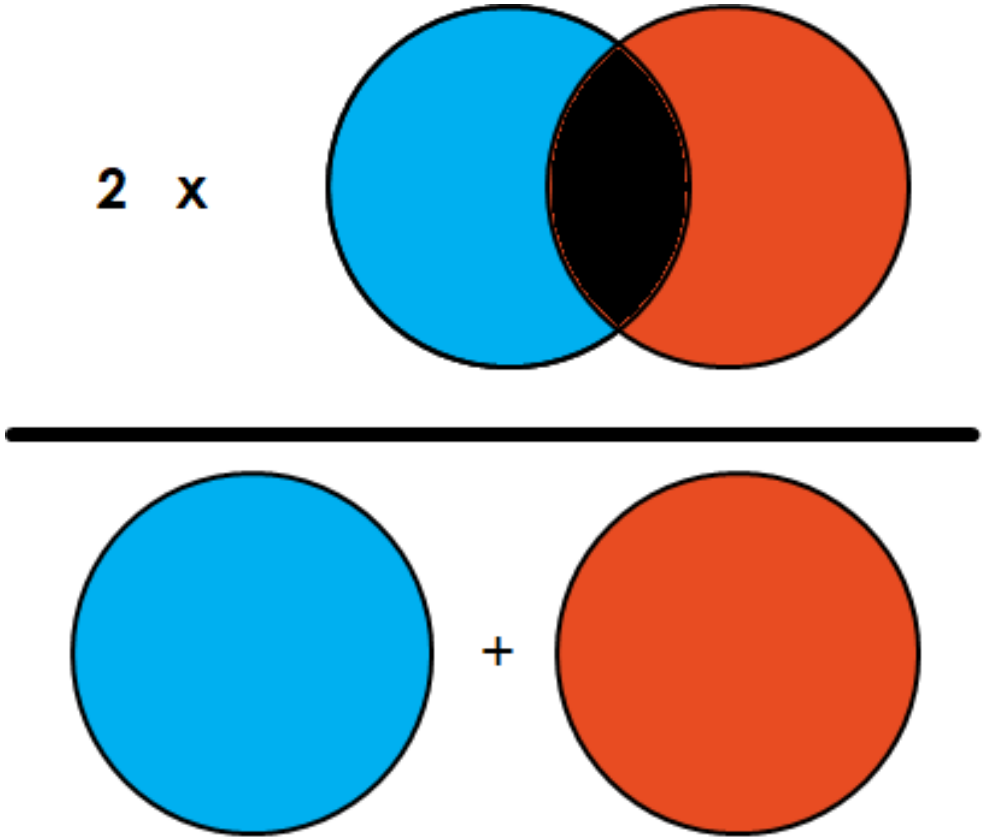


Gitonga, M. M. (2023). *Multiclass MRI Brain Tumor Segmentation using 3D Attention-based U-Net*. arXiv preprint arXiv:2305.06203. https://arxiv.org/abs/2305.06203

DICE Coefficient



$$\frac{2 * |X \cap Y|}{|X| + |Y|}$$



Workshop 4 – Tumor Semantic Segmentation

https://colab.research.google.com/drive/1F5xrndrCGmRqiyZbsMDymTnrrHXWjcV8?usp=sharing

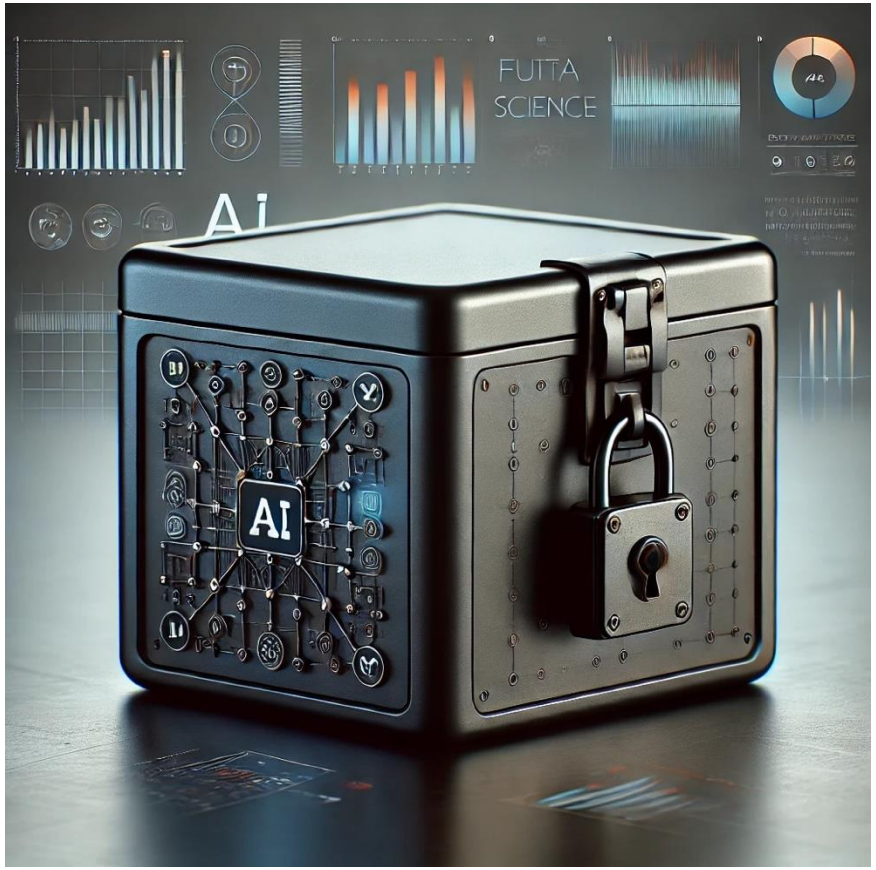
MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

ihealth | Millennium Institute for Intelligent Healthcare Engineering

The Black Box Problem

Universidad de Valparaíso CHILE

The Black Box Problem of Artificial Intelligence



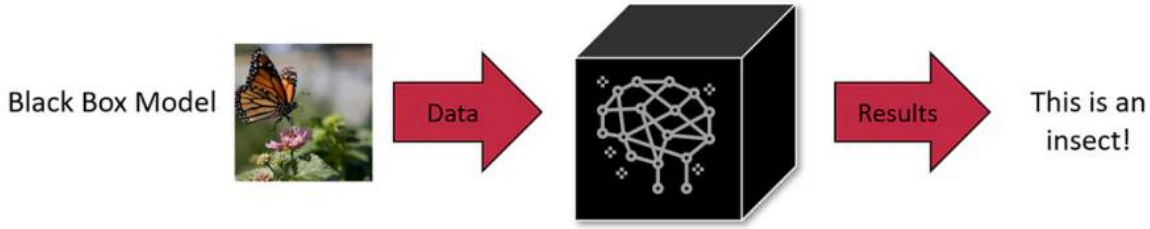
The Black Box Problem in Artificial Intelligence

Lack of Transparency
No clear explanation of how predictions are made.

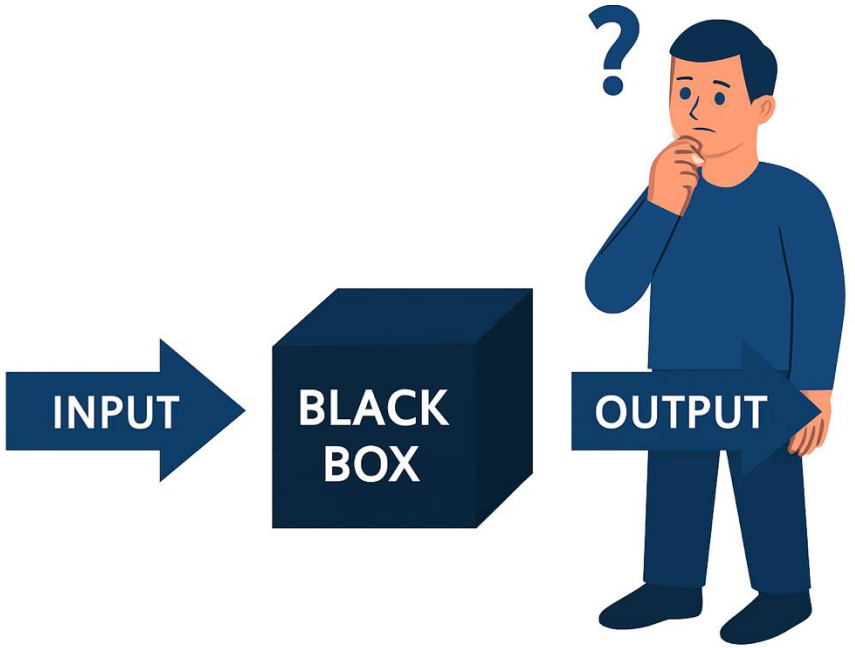
Limited Trust
Difficult for medical professionals to rely on outputs they cannot understand.

Bias and Hidden Errors
Potentially dangerous if incorrect decisions go undetected.

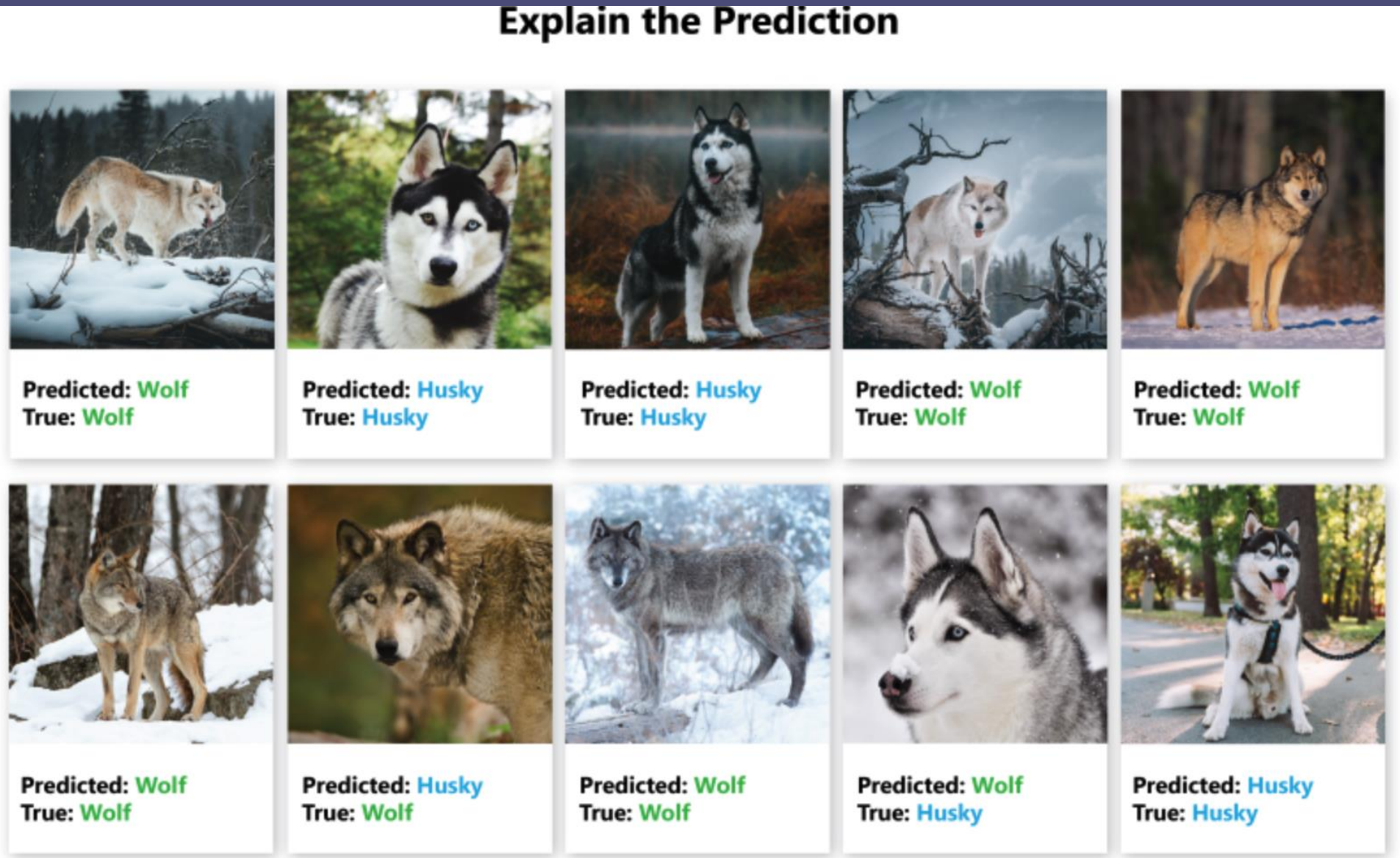
Ethical Concerns
Users need understandable reasoning behind critical decisions.



https://www.unite.ai/the-black-box-problem-in-llms-challenges-and-emerging-solutions/



Huskies vs Wolf



Besse, Philippe & Castets-Renard, Céline & Garivier, Aurélien & Loubes, Jean-Michel. (2018). Can Everyday AI Be Ethical? Machine Learning Algorithm Fairness (English version). 10.13140/RG.2.2.22973.31207.

Huskies vs Wolf

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

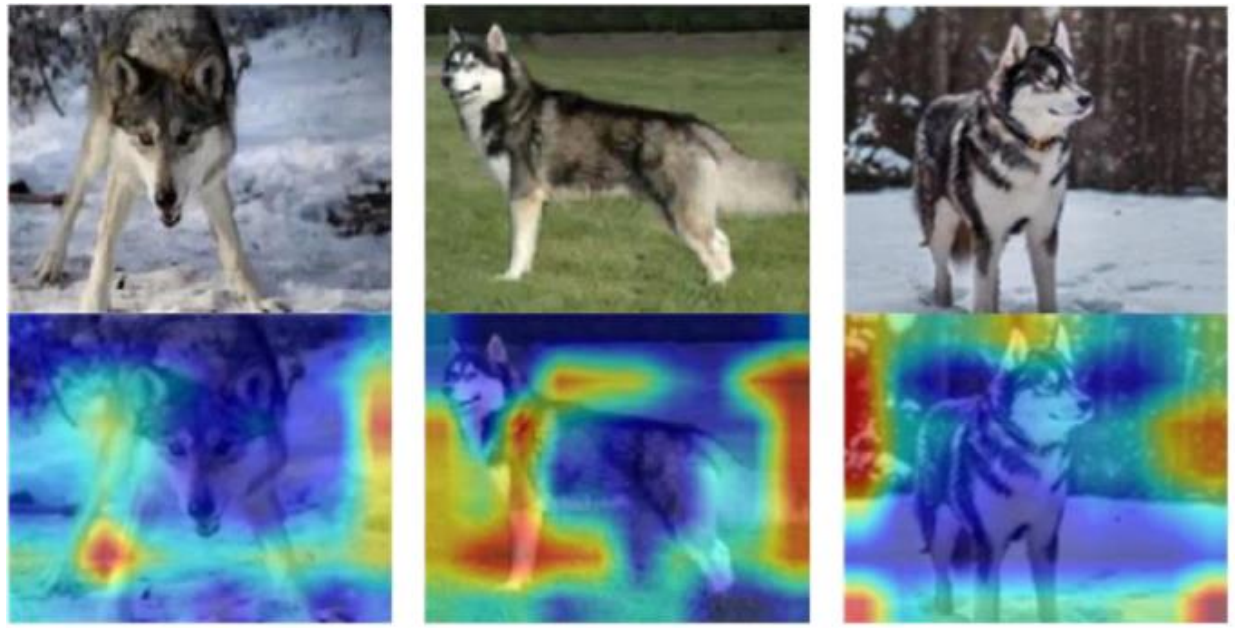
Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.



(a) The image of the Arctic Wolf in the background of snow is correctly classified.

(b) The image of the Husky in the background of common living areas is correctly classified.

(c) The image of the Husky in the snow is classified as "Arctic Wolf".

¡Open the Black Box!

MEDING
Interdisciplinary Center for Biomedical and Health Engineering Research

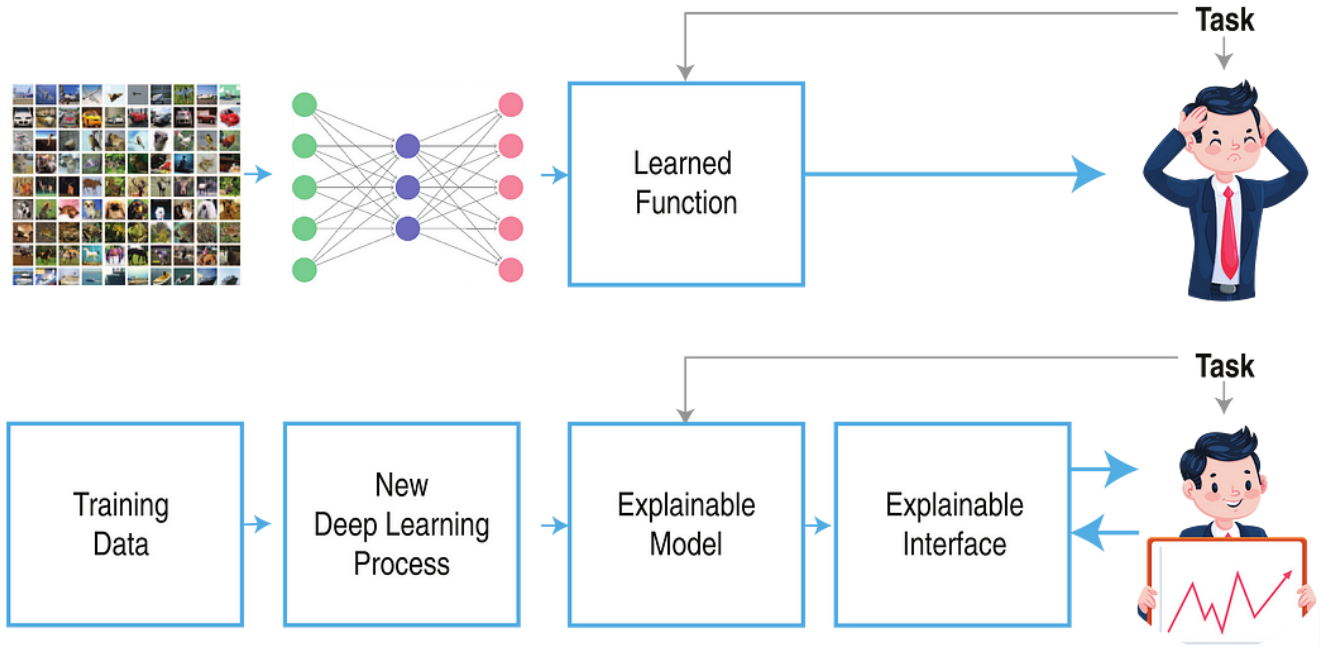
ihealth | Millennium Institute for Intelligent Healthcare Engineering

eXplainable Artificial Intelligence

XAI

Universidad de Valparaíso CHILE

What is Explainable Artificial Intelligence



https://medium.com/deepviz/what-is-xai-explainable-ai-and-visualization-part-10-da41c981c5fa

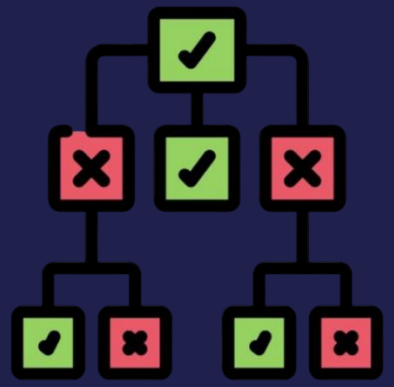
Benefits of Explainable AI

- Enables the deployment of trustworthy and understandable AI models.
- Enables continuous evaluation for faster AI outcomes.
- Reduces risk, bias, and auditing costs in AI systems.

Explainable Artificial Intelligence(XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by AI algorithms.

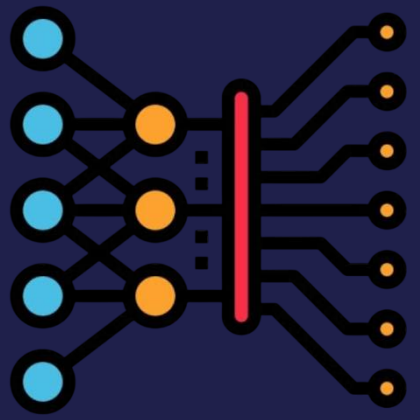
Interpretability vs Explainability

INTERPRETABILITY



Describes *how* the model works internally, from the **developer's perspective**.

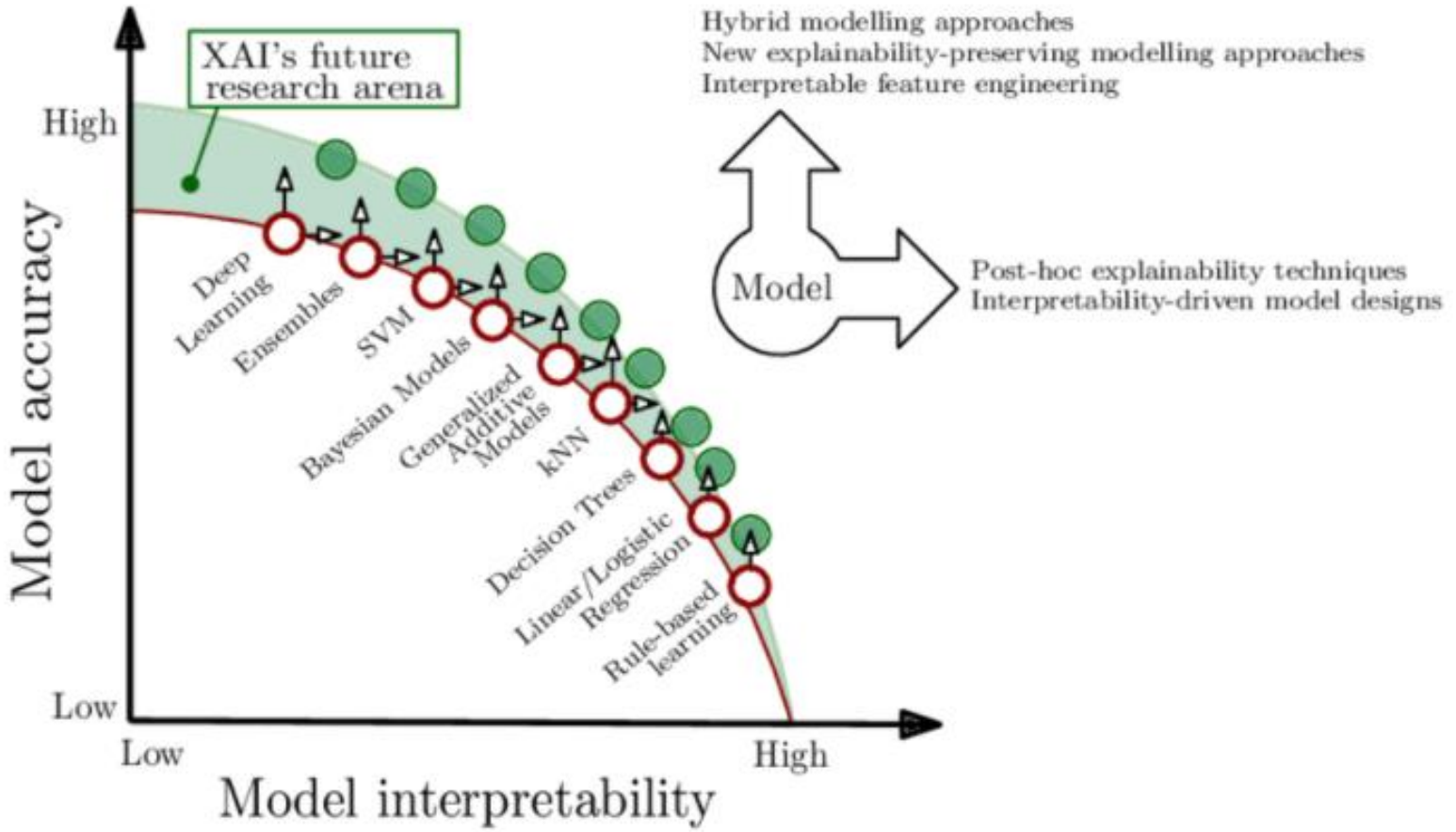
EXPLAINABILITY



Explains *why* the model made a decision, from the **user's perspective**.

https://medium.com/data-science/interperable-vs-explainable-machine-learning-1fa525e12f48

Open the Black Box



https://doi.org/10.1016/j.coelec.2024.101629

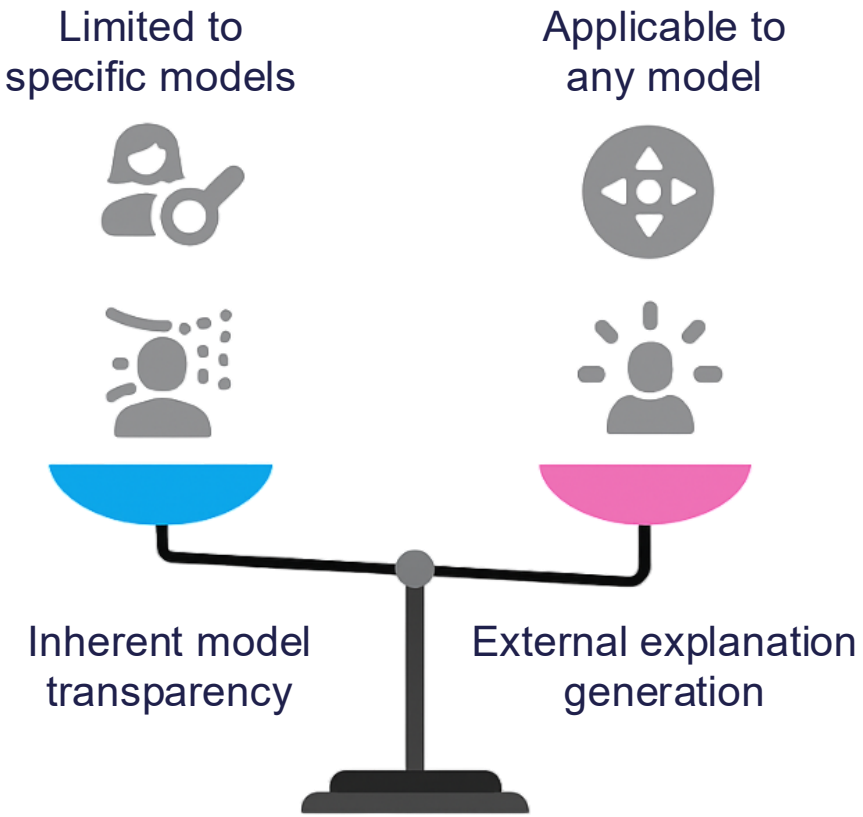
https://doi.org/10.1016/j.cirp.2024.06.003

Post-Hoc vs. Ante-Hoc in XAI

Refers to models designed to be explainable from the start.

ANTE-HOC

Their internal structure allows for understanding without additional techniques.



Applies explanatory methods after training complex models.

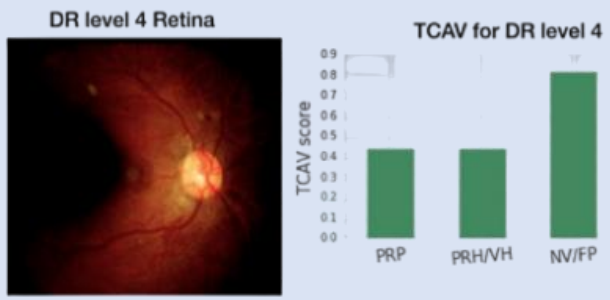
POST-HOC

Aims to justify or interpret decisions made by black-box models.

https://doi.org/10.1016/j.cogsys.2024.101243

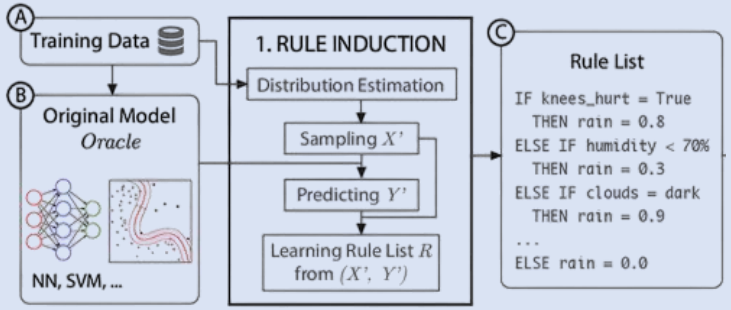
Post-Hoc Explainability Approaches

https://doi.org/10.48550/arXiv.1711.11279



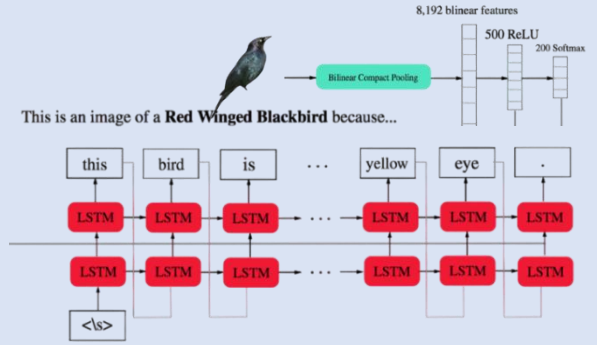
Numeric Explanations

http://dx.doi.org/10.1109/TVCG.2018.2864812

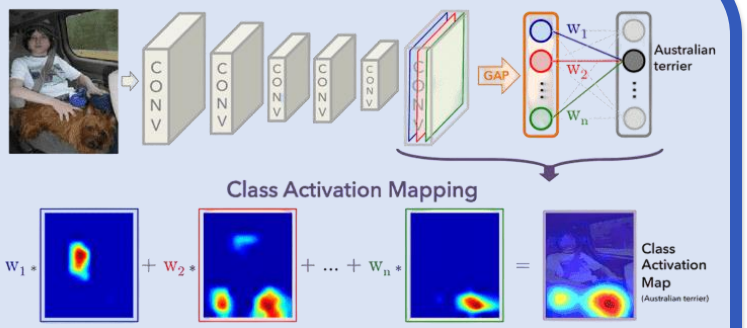


Rule-Based Explanations

https://doi.org/10.3390/make3030032

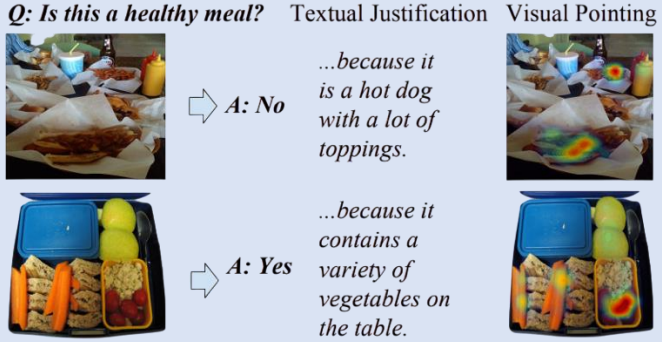


Textual Explanations



Visual Explanations

https://doi.org/10.48550/arXiv.1512.04150



Mixed Explanations

https://doi.org/10.3390/make3040048

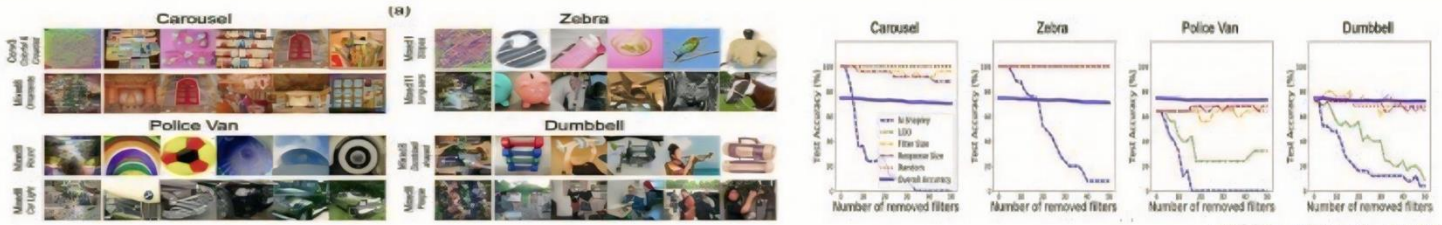
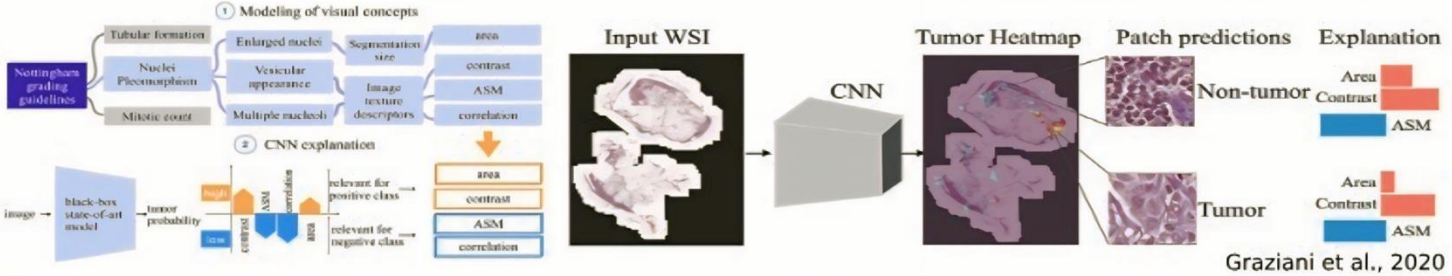
Visual Explainability in Deep Learning

MSc. Student
Esthefanía Astargo

Material desarrollado con el apoyo de la estudiante del Magister en Ciencias e Ingeniería para la Salud, Universidad de Valparaíso
Sr. Esthefanía Astargo

XAI in Images

Image

Saliency Map Visualization	 Simonyan et al., 2013
Shapley Value Importance	 Ghorbani et al., 2020
Concept Attribution	 Graziani et al., 2020

https://spectra.mathpix.com/article/2021.09.00007/demystify-post-hoc-explainability

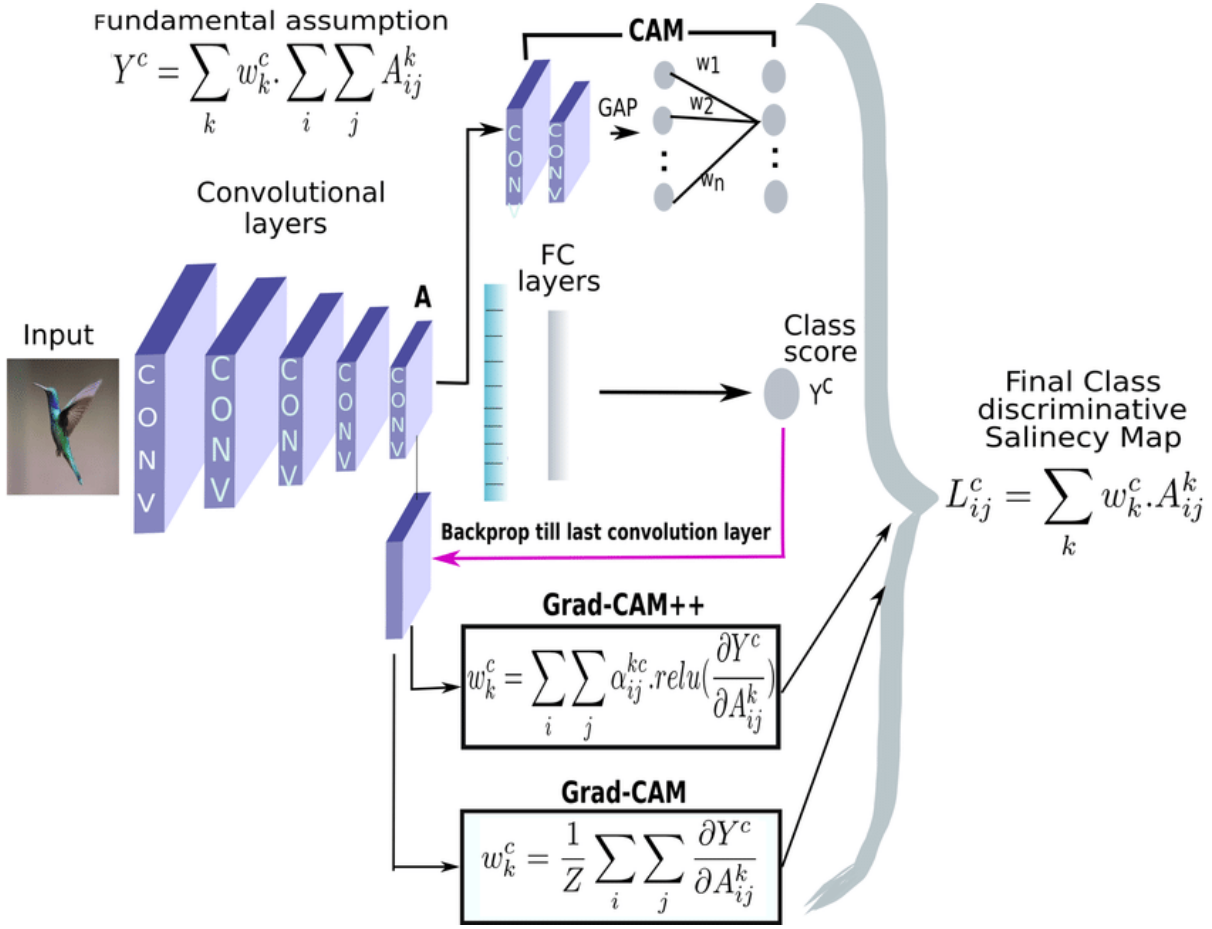
Grad-CAM: Gradient-weighted Class Activation Mapping

Grad-CAM es un método visual post-hoc y específico de modelo, diseñado para redes convolucionales profundas (CNNs).

Se basa en los gradientes del modelo para identificar qué regiones del input activaron más una clase específica.

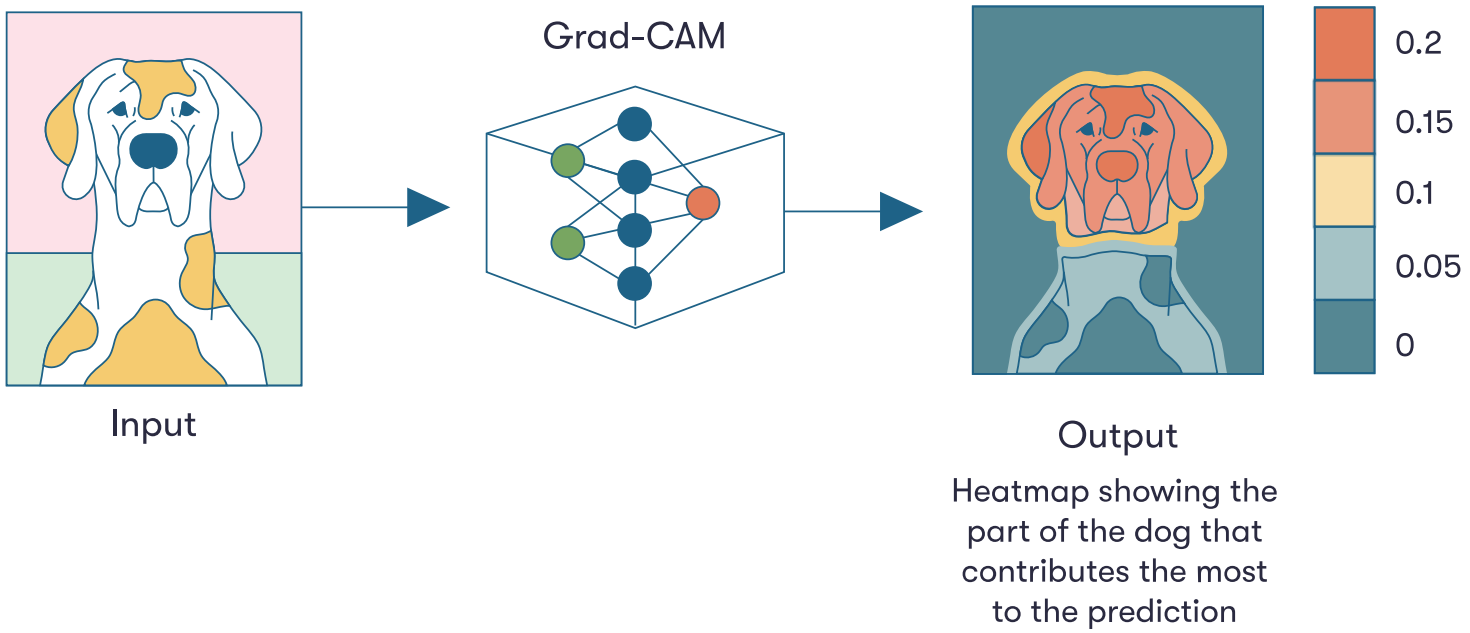
Utiliza los mapas de activación de la última capa convolucional, conservando la información espacial relevante.

Calcula un mapa ponderado de relevancia combinando gradientes y activaciones, destacando zonas clave.



https://doi.org/10.48550/arXiv.1710.11063

Grad-CAM: Gradient-weighted Class Activation Mapping

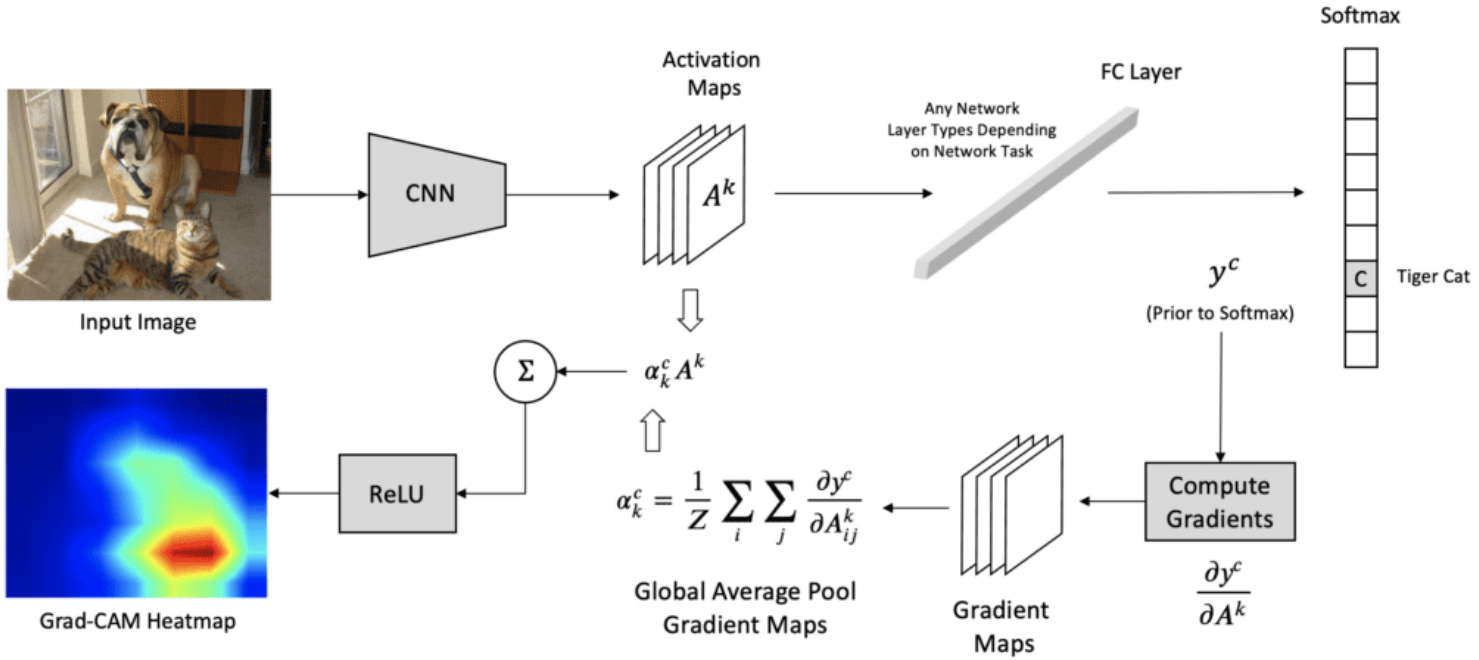


https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/explainability/types-of-explainable-ai/

Grad-CAM aplicado a clasificación de imágenes

- El modelo recibe como entrada la imagen de un perro y realiza una predicción de clase.
- Grad-CAM propaga hacia atrás los gradientes desde la salida hasta la última capa convolucional.
- A partir de los mapas de activación y gradientes, se genera un mapa de calor.
- El heatmap superpuesto indica qué regiones del perro fueron más relevantes para la predicción.

Grad-CAM paso a paso



https://xai--tutorials-readthedocs-io.translate.goog/en/latest/_model_specific_xai/Grad-CAM.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Paso 1: Forward Pass

- Obtener los mapas de características de la última capa convolucional y las salidas sin procesar antes de la función softmax.
- Estos mapas de características se denominan A^k, donde k es un mapa de características específico dentro de una capa convolucional.

Paso 2: Selección de la clase objetivo

- Se elige la clase **c** que se desea explicar (normalmente la clase predicha con la puntuación más alta) y se toma su valor de activación antes de la función softmax.
- y^c representa la activación asociada a la clase c antes de aplicar la función softmax.

Grad-CAM paso a paso

Paso 3: Calcular los gradientes

- Se calculan los gradientes de y^c con respecto a los mapas de características A^k, es decir:

$$\frac{\partial y^c}{\partial A^k}$$

Paso 4: Calcular el mapa Grad-CAM

- Se calcula un peso por cada mapa (media global de gradientes):

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

- Luego se combinan todos los mapas con sus pesos:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

- La función ReLU conserva solo las regiones que contribuyen positivamente.

https://xai--tutorials-readthedocs-io.translate.goog/en/latest/_model_specific_xai/Grad-CAM.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

Workshop 5 – Using Grad-CAM for Explainability in Image Classification

Google colab

https://colab.research.google.com/drive/1qIAtUqSFklVSK2LWjMBzvVM4nGAC1CvO?usp=sharing

LIME: Local Interpretable Model-agnostic Explanations

Construye un modelo sustituto que aproxima al modelo complejo solo en torno a una predicción específica.

Genera variaciones del dato de entrada y observa cómo el modelo original responde a estos cambios.

Utiliza esas observaciones para entrenar un modelo simple que imita al modelo complejo localmente.

Este modelo sustituto revela qué características influyen más en esa predicción puntual.

Puede aplicarse a texto, imágenes, o datos tabulares, sin requerir acceso interno al modelo.

Matemáticamente, los modelos sustitutos se pueden expresar de la siguiente manera:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

f es el modelo complejo original.

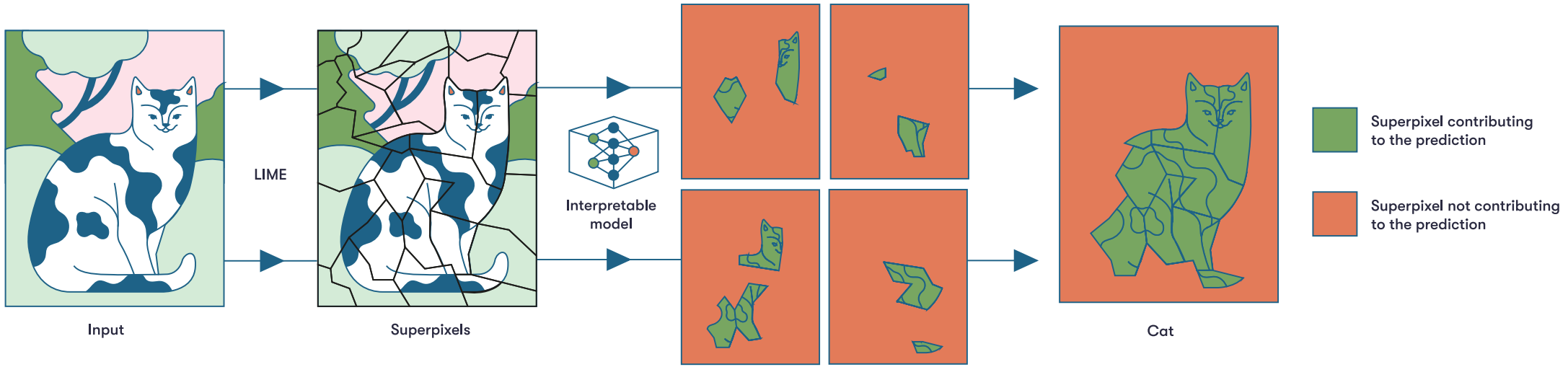
g es el modelo interpretable local.

$\mathcal{L}(f, g, \pi_x)$ mide cuán bien g aproxima a f cerca de x.

π_x define la vecindad local ponderando las instancias perturbadas.

$\Omega(g)$ penaliza la complejidad de g.

LIME: Local Interpretable Model-agnostic Explanations



https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/explainability/types-of-explainable-ai/

LIME aplicado a clasificación de imágenes

- Divide la imagen en superpíxeles, que agrupan regiones visuales coherentes y comprensibles.
- Se generan múltiples versiones de la imagen desactivando distintas combinaciones de superpíxeles
- El modelo original predice sobre cada versión modificada y LIME observa cómo cambian las salidas.
- Luego, entrena un modelo interpretable para estimar qué superpíxeles son claves en la decisión.
- Ejemplo: los superpíxeles verdes explican por qué el modelo clasificó la imagen como *gato*.

Workshop 6 - Uso de LIME para Explicabilidad en Clasificación de Imágenes

Google colab

¡Escanéame!

https://colab.research.google.com/drive/1pWRDK32SPgBA3OD83_8qkTwsuBBkqqaz?usp=sharing

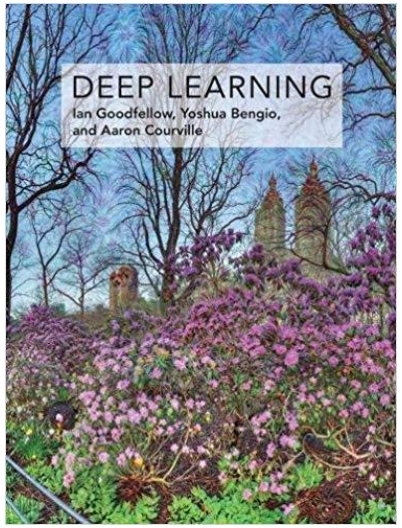
Material del Curso y la Conferencia

Github

https://github.com/rodsalasf/

Dr. Rodrigo Salas Fuentes
rodrigo.salas@uv.cl

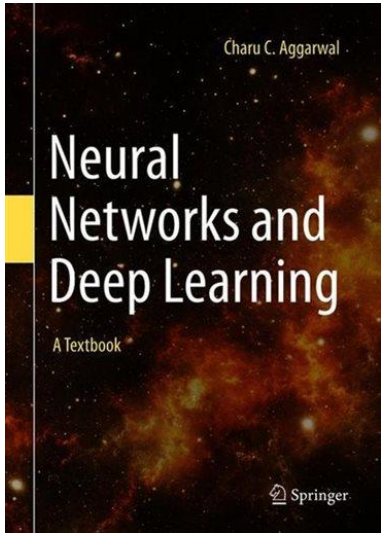
Deep Learning Books



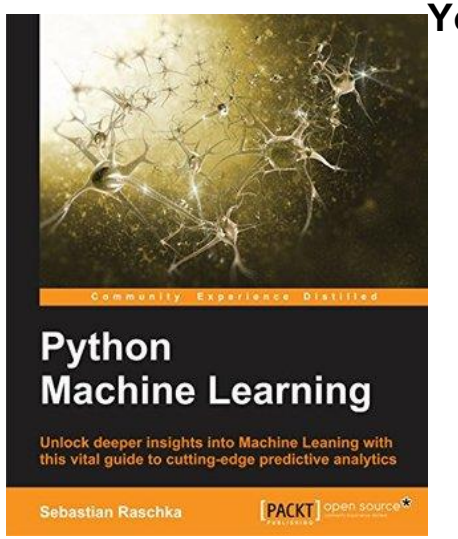
Ian Goodfellow

Yoshua Bengio

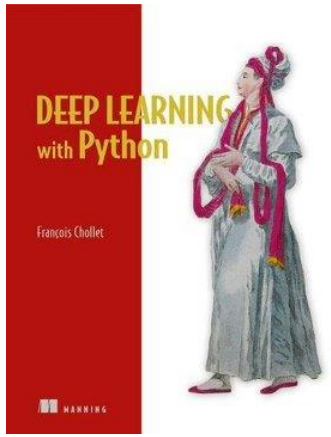
Aaron Courville



Charu Aggarwal



Sebastian Raschka



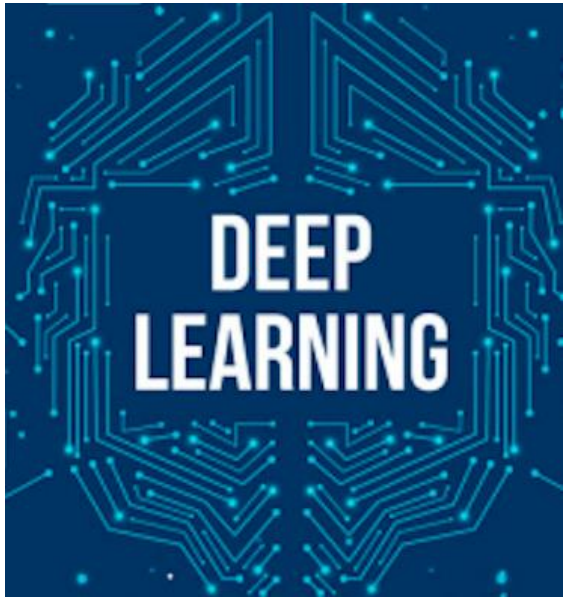
Francois Chollet

Toolbox for Deep Learning

torch

PYTORCH
Deep Learning with PyTorch

Caffe



TensorFlow

Microsoft CNTK

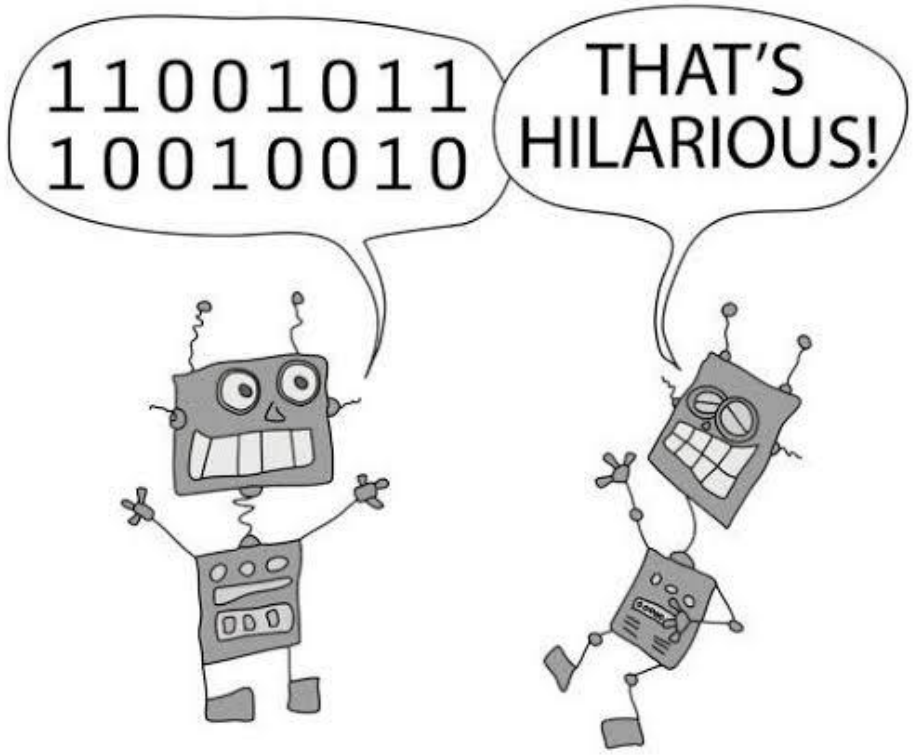
Keras
A deep learning library

theano

Caffe2

dmlc mxnet


Preguntas?
Grad-CAM
XAI



Dr. Rodrigo Salas Fuentes
rodrigo.salas@uv.cl

Thanks for your Attention

LinkedIn: linkedin.com/in/rodrigo-salas-fuentes/
Google Scholar: scholar.google.com/citations?user=ZaqDIPcAAAAJ
ORCID: orcid.org/0000-0002-0350-6811
Email: rodrigo.salas@uv.cl

Ingeniería Civil Biomédica
Facultad de Ingeniería
Universidad de Valparaíso
www.biomedica.uv.cl