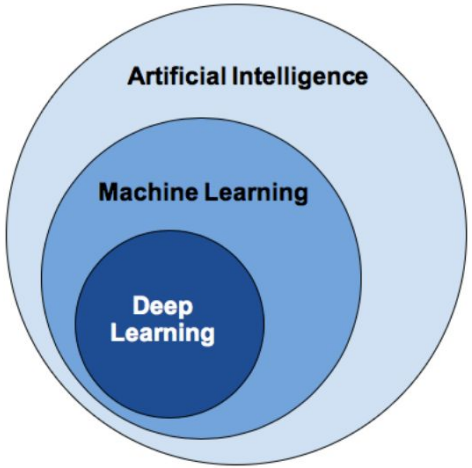
Deep Learning
Procesamiento de video - Reconocimiento de LS

Escuela de Informática - CACIC 2024
Lic. Pedro Dal Bianco - dalbianco.pedro@gmail.com

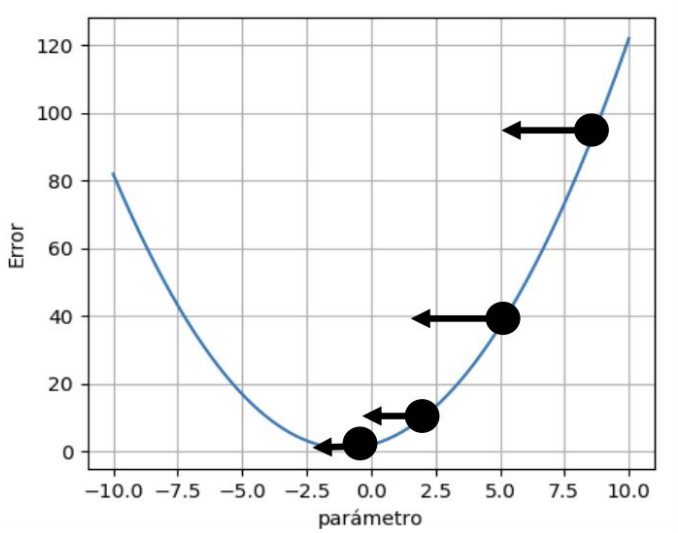
Deep Learning

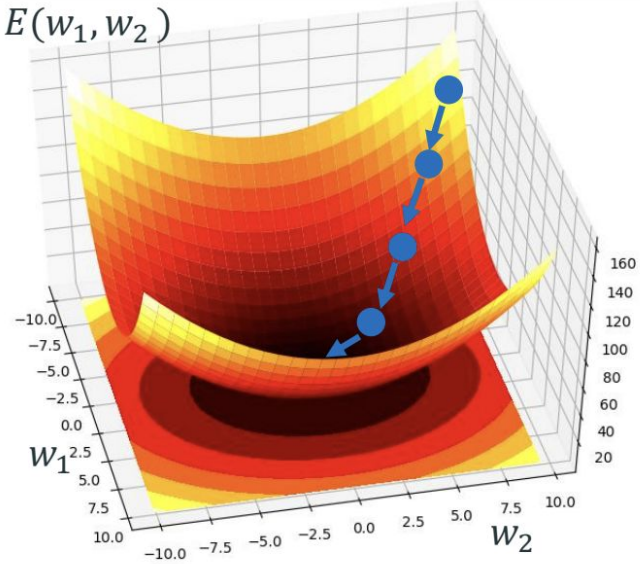
IA, ML y Deep Learning

- **AI**: *"Disciplina que se ocupa de crear programas que ejecutan operaciones comparables a la mente humana."*
- **ML**: *"estudia mecanismos que permite a las computadoras aprender a clasificar o predecir en base a experiencias."*
- **DL**: *"machine learning utilizando grandes modelos neuronales y cantidades de datos"*

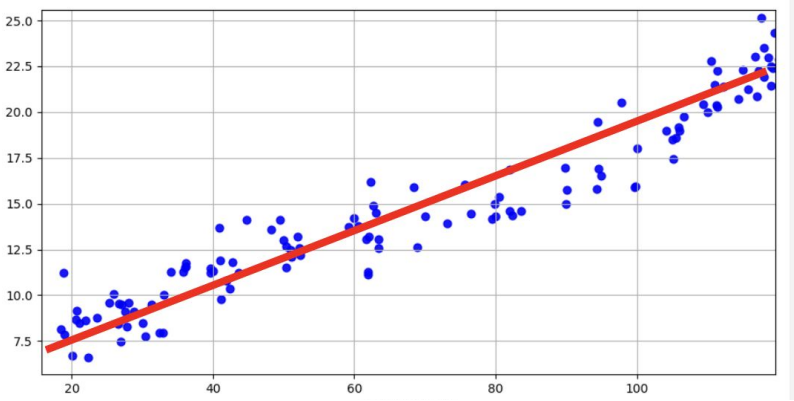


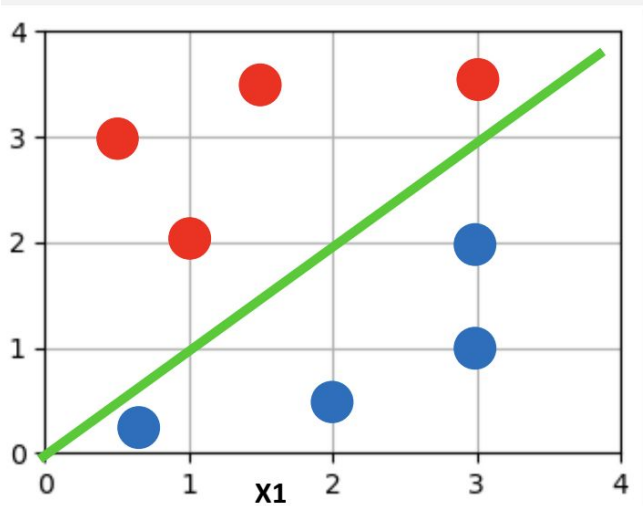
Descenso de gradiente



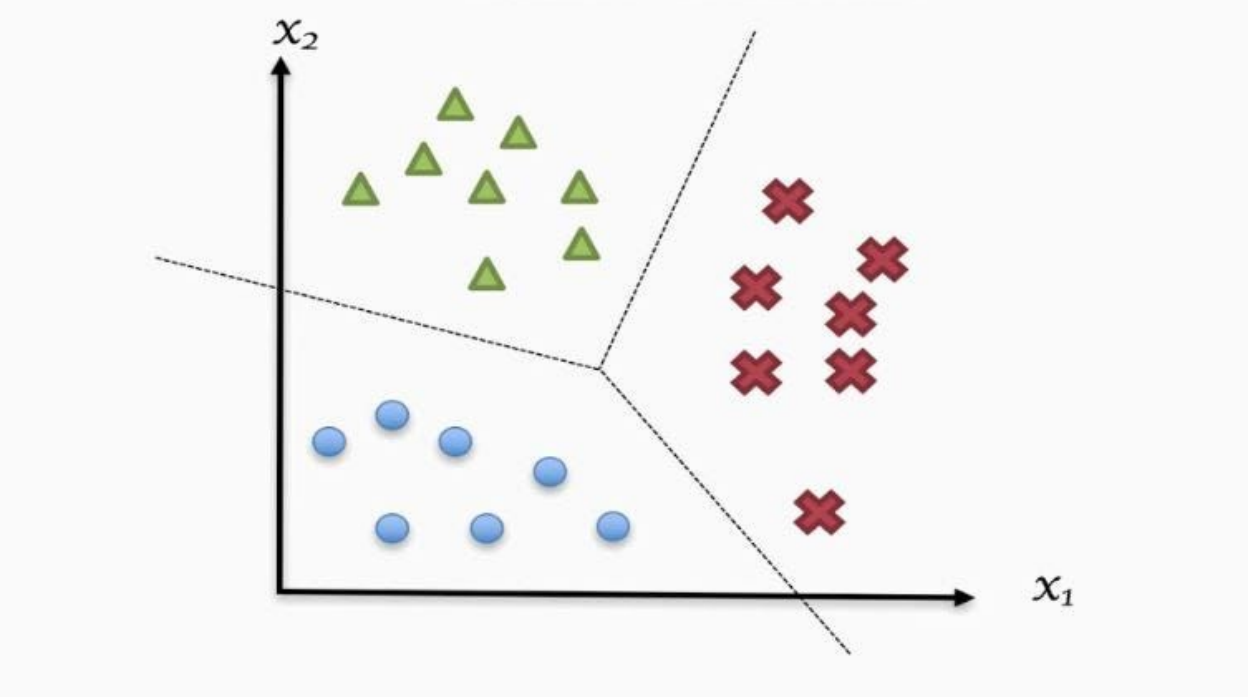


Deep Learning - Regresión / Clasificación

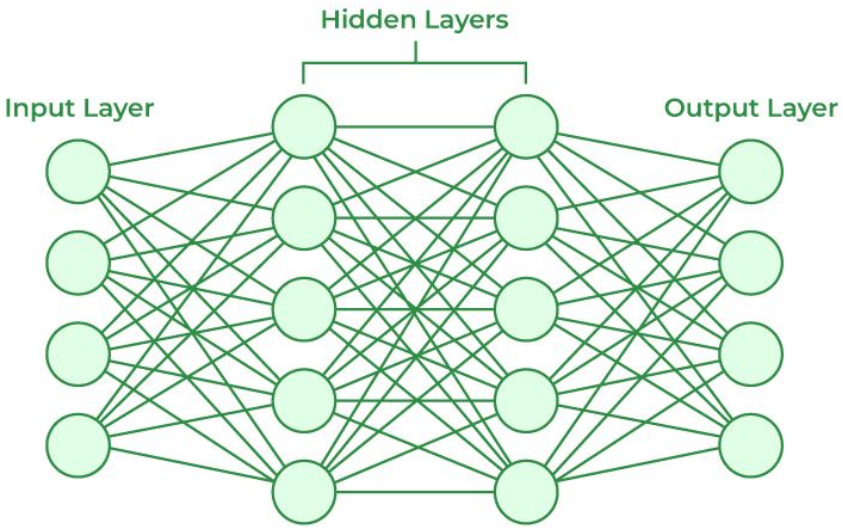




Deep Learning - Clasificación Multiclase

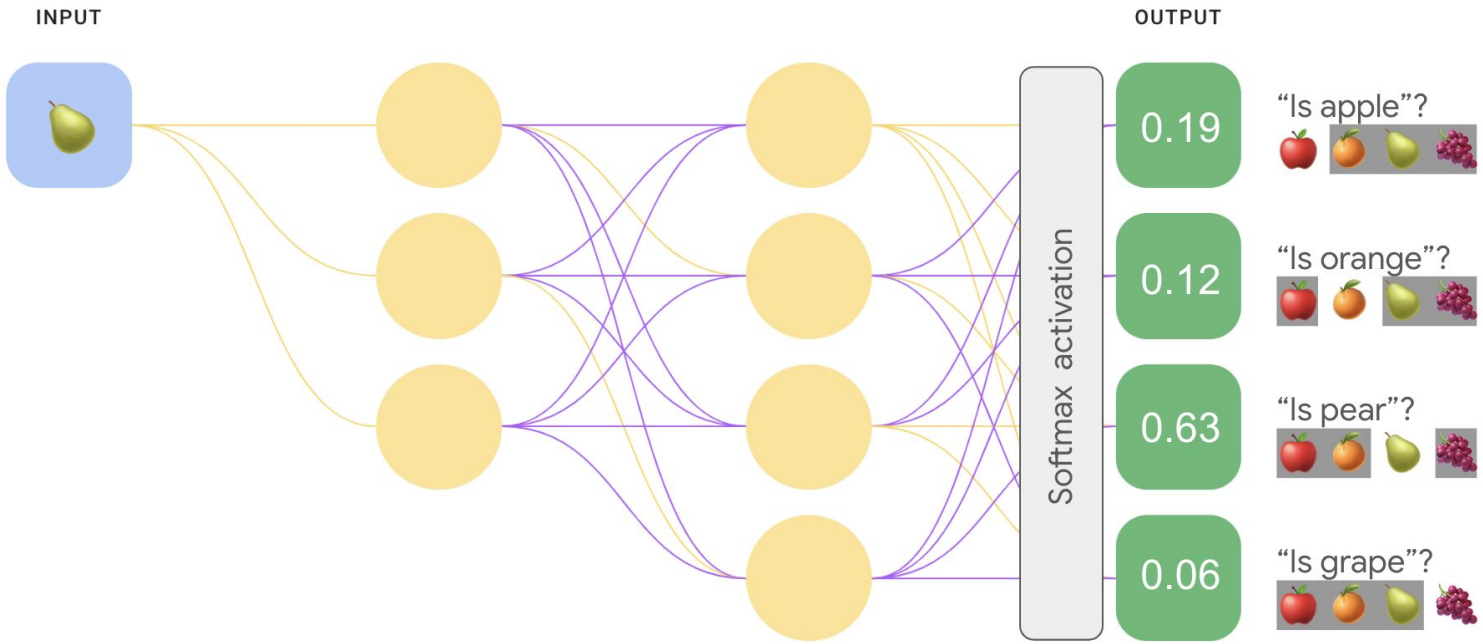


Redes neuronales



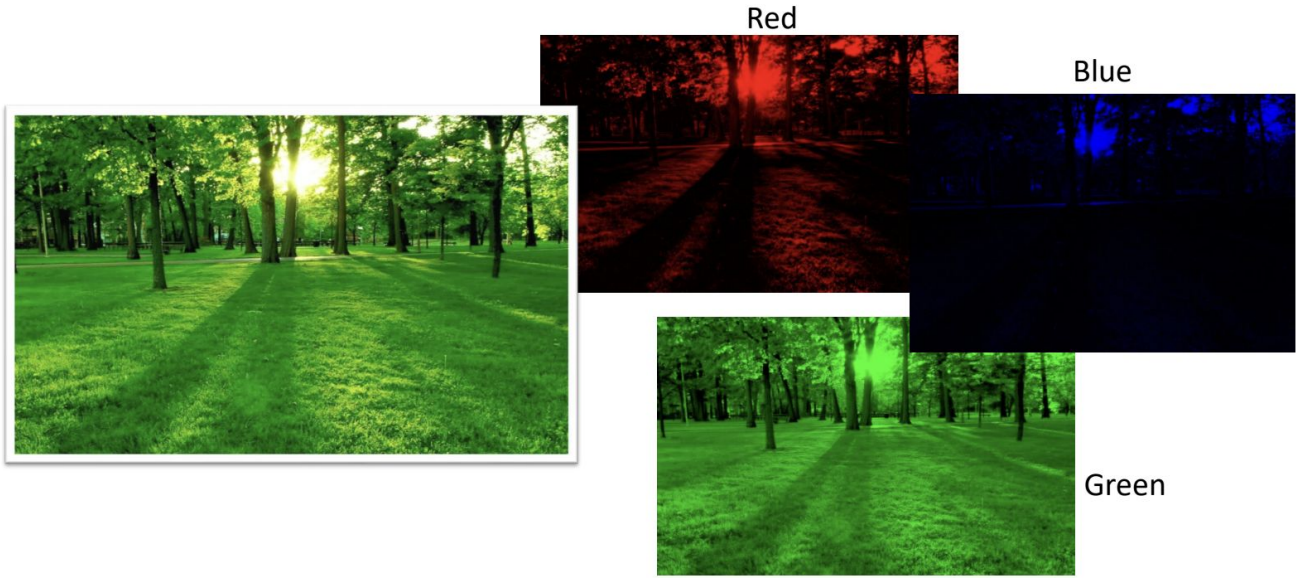
Aproximadores universales: http://neuralnetworksanddeeplearning.com/chap4.html

Redes neuronales



Imágenes

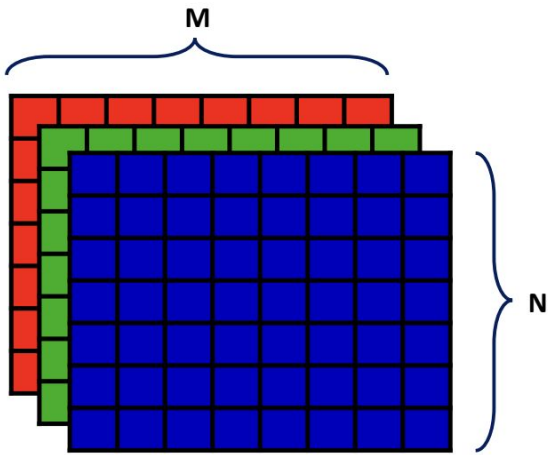
Representación de imágenes



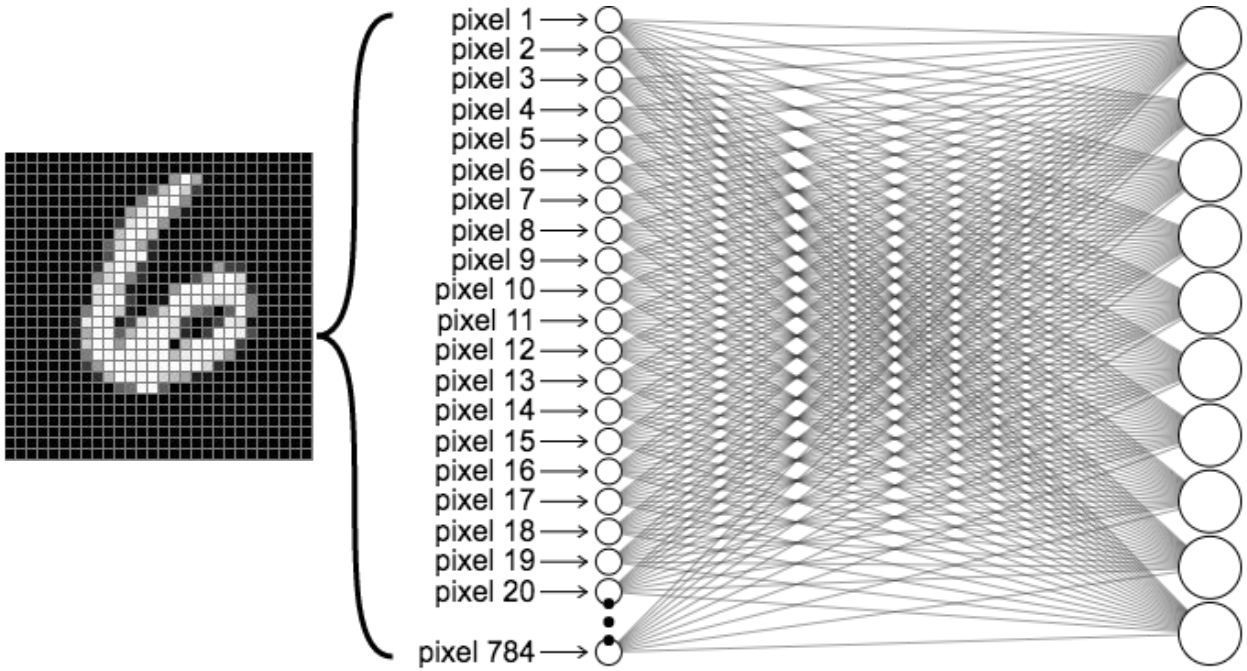
Representación de imágenes

Una imagen color se representa con 3 matrices de NxM (una por canal).

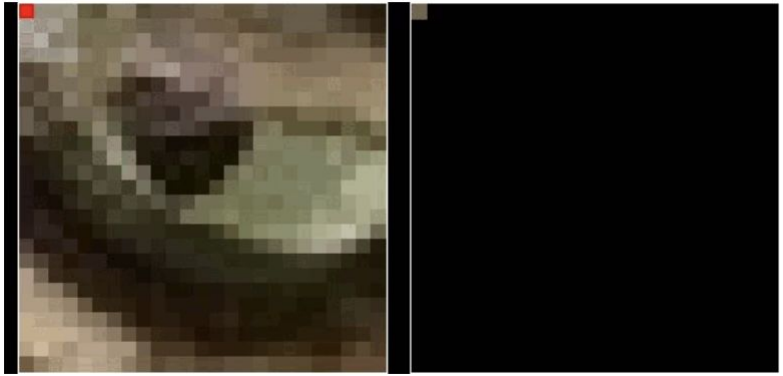
- Imagen full HD: 1080 x 1920 x 3 = ***6.220.800 features***

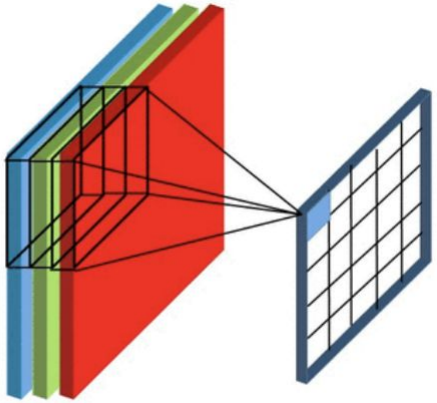


Imágenes en FF

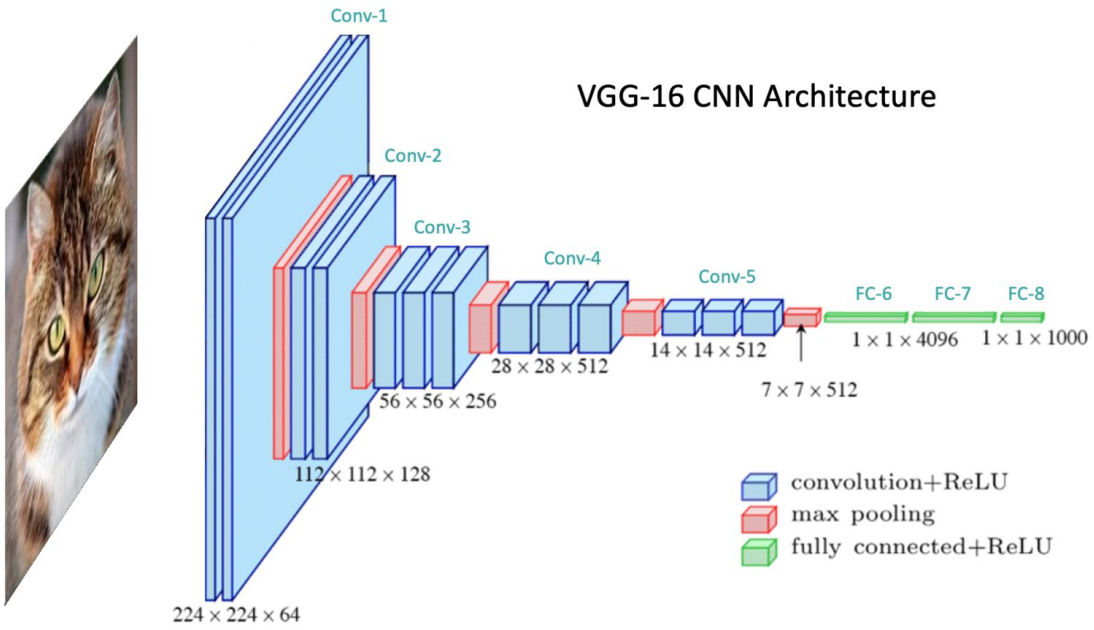


Convoluciones



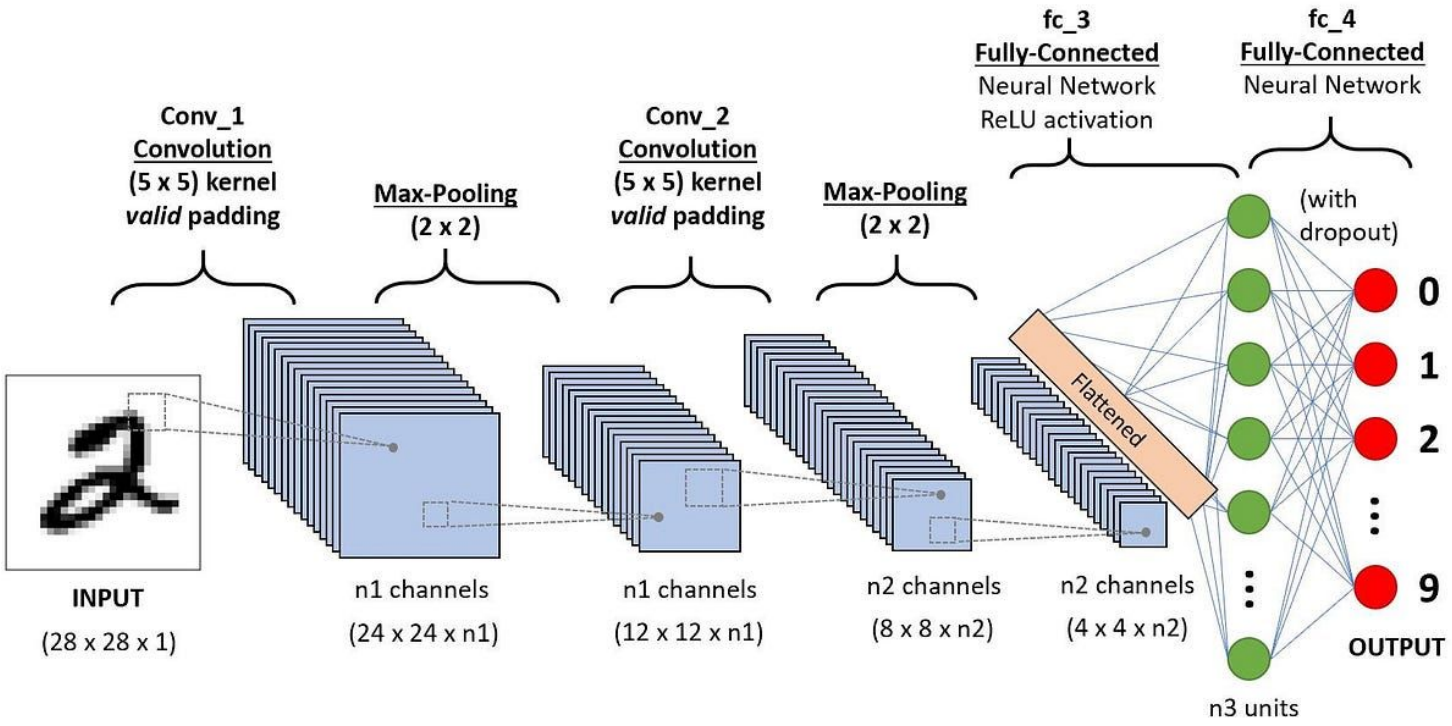


CNN



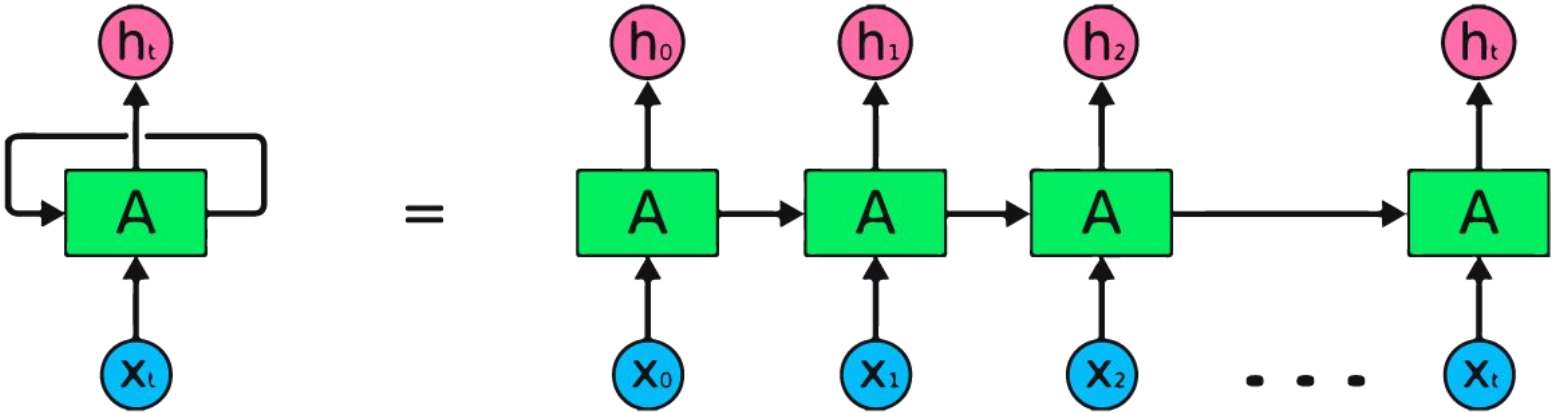
Ejemplo: https://tensorspace.org/html/playground/lenet.html

Clasificación con imágenes

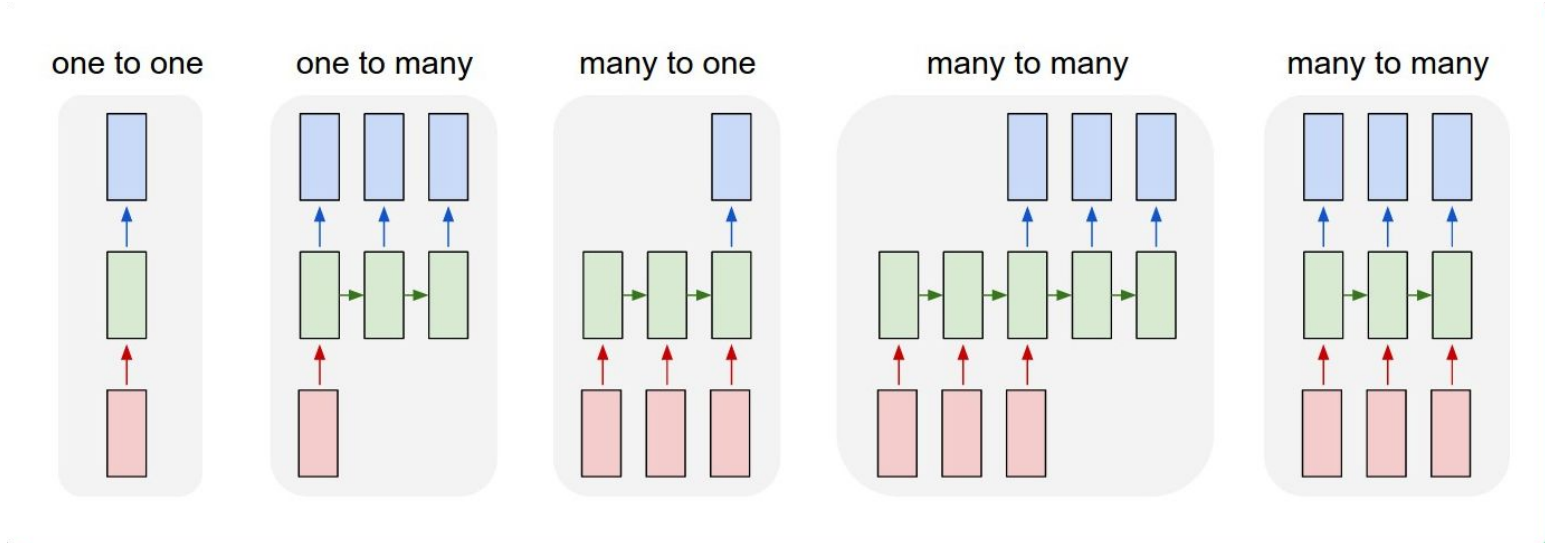


Representaciones temporales

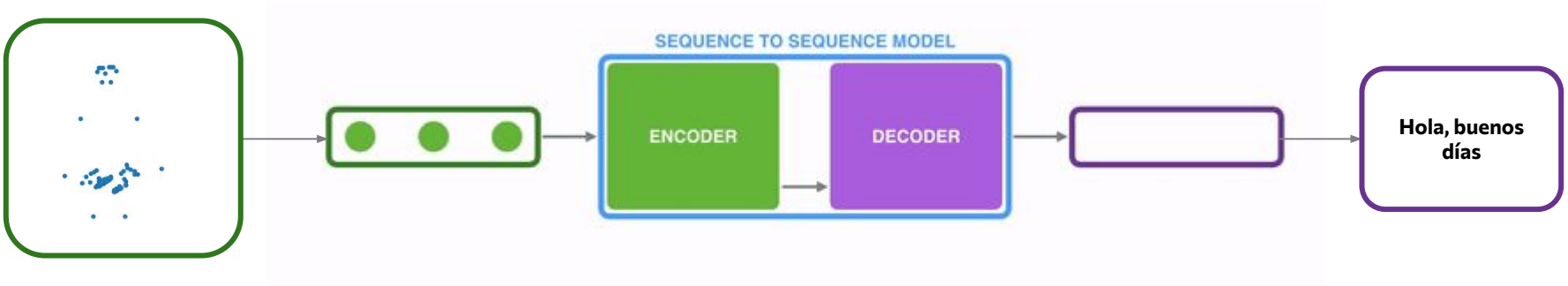
RNN



RNN para clasificación



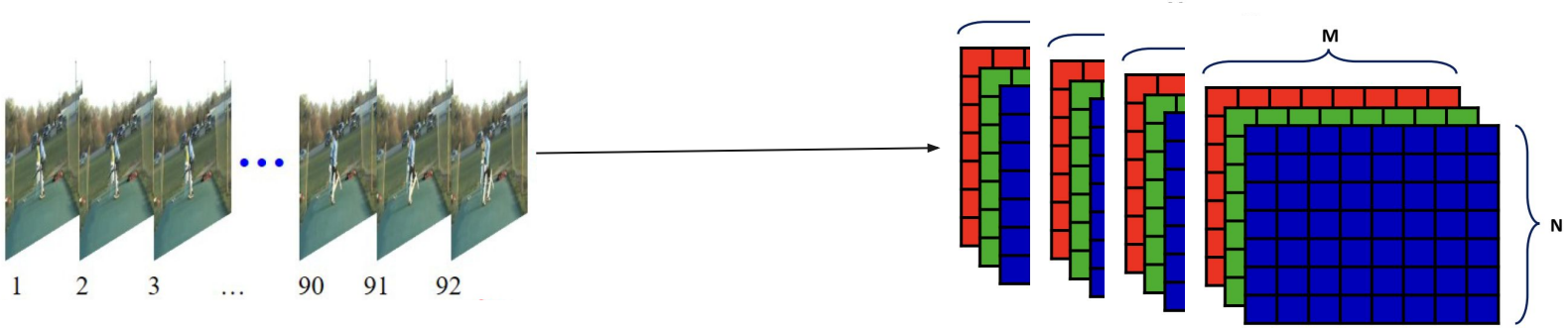
Redes Encoder-Decoder



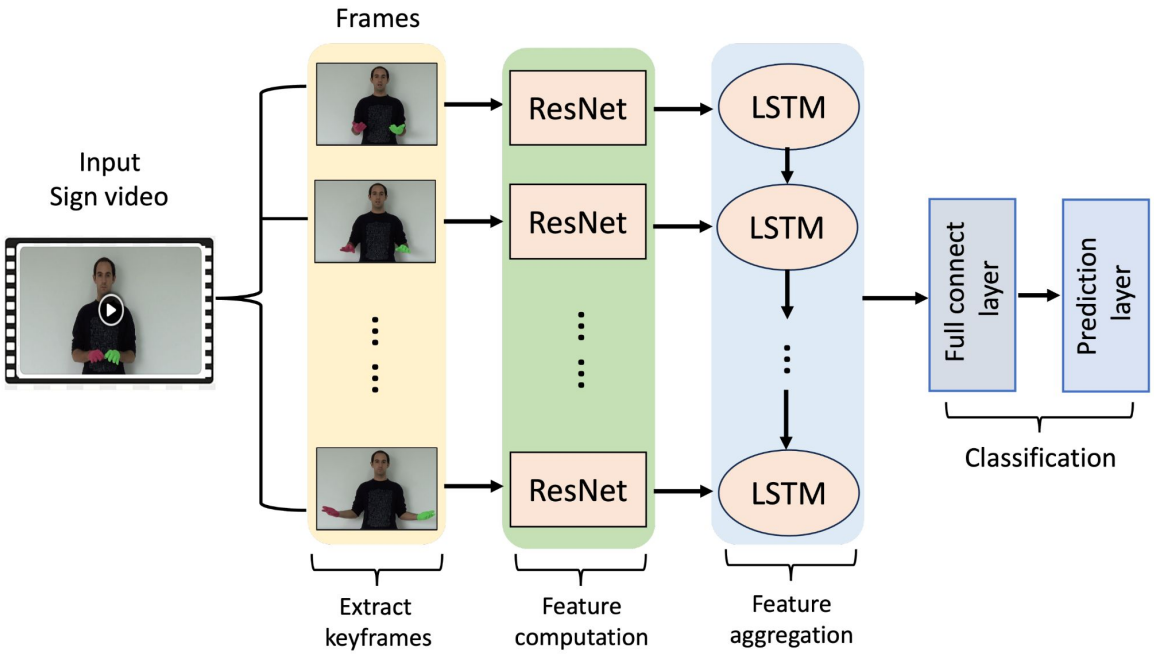
Reconocimiento de LS

Representación de video

Si una imagen en full HD se representaba con ***6 millones de features***, en un video esto se multiplica por la cantidad de frames



Modelo de SLR



LSA 64

- Desarrollado en 2016 por III-LIDI, UNLP
- Contiene 64 señas realizadas por distintos señantes repetidas veces.

Traducción de LS

Reconocimiento de Lengua de Señas

- **Reconocimiento de Lengua de Señas**: clasificar un video -> una seña

SLR Model

“Llamar”

Traducción de Lengua de Señas (SLT)

SLT: Video -> Traducción a idioma escrito

- No hay mapeo uno a uno entre seña y palabras
- No comparten reglas gramaticales

SLT Model

“ Bienvenidos a CN Sordos Argentina, el primer canal de noticias en LSA realizado por personas sordas”

Desafíos

Regionalización

“Hola” en tres lenguas de señas distintas

Argentina

EEUU

China

Segmentación temporal

Señas aisladas no son suficientes

“Llamar”

“Bienvenidos a CN sordos, el primer noticiero en Lengua de Señas Argentina”

Gestos multicanal

Involucran la forma y posición de manos, movimientos, posturas corporales y expresiones faciales

Falta de datos

Dataset	Formato	Cantidad
ImageNet	Imágenes	14 millones
CommonCrawl	Tokens (~palabras)	410.000 millones
LibriSpeech ASR	Audio y transcripción.	150.000 (~1000 hs)
RWTH-Phoenix	Video y traducción	7.000 (~8hs)

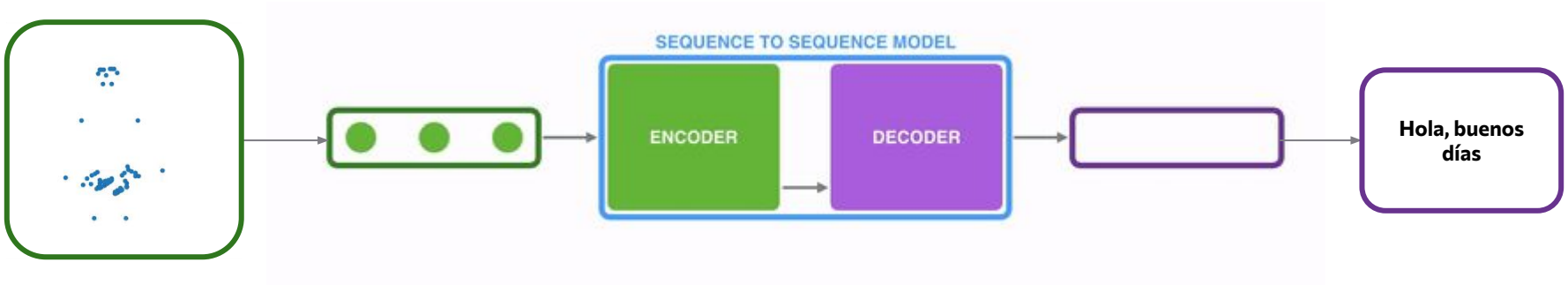
Sesgos en los datasets

Poca variedad de señantes, fondos, iluminación y temas

Modelos de SLT

Redes Encoder-Decoder

- Modelos de lenguaje del tipo encoder decoder
- A partir de el video como entrada, predicen de a una la siguiente palabra



RWTH Phoenix

- Principal benchmark para modelos de SLT
- Creada a partir de reportes del clima de la televisión alemana
- Poca variedad de señantes, ropa, fondo e iluminación

BBDD LSA-T

- Primer dataset para traducción de LSA a partir del canal CN Sordos.
- Diversidad de temas, señantes, escenarios, vestimenta e iluminación
- Contiene información posicional

Actividad

Modelo de SLR sobre LSA64

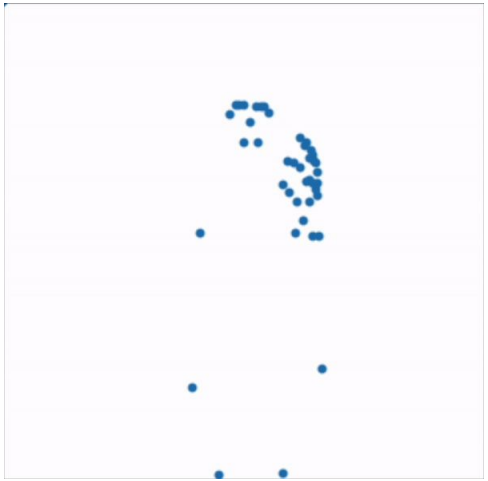
LSA 64

- Desarrollado en 2016 por III-LIDI, UNLP
- Contiene 64 señas realizadas por distintos señantes repetidas veces.

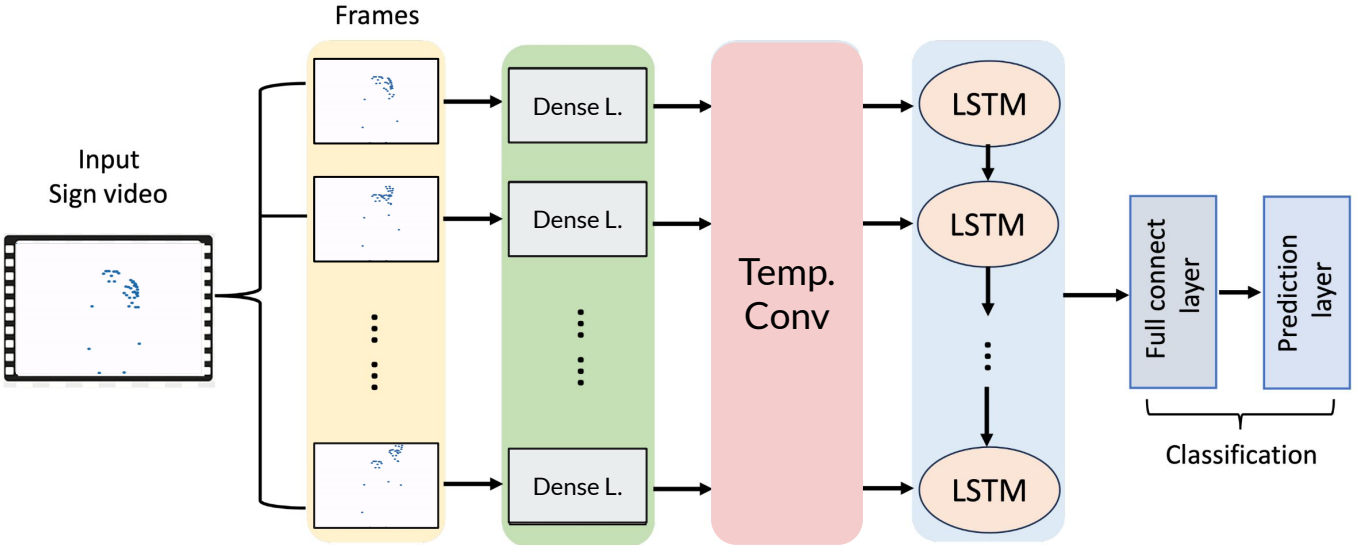
Extracción de poses

Se utilizó MediaPipe para extraer las poses

Un frame full HD tiene aprox 6M de features, un frame de info. posicional 150



Modelo de SLR



Herramientas

- Se utiliza **Keras** como framework para construir el modelo neuronal

- Se utiliza **Kaggle** para hostear la competición, almacenar el dataset, ejecutar los modelos y evaluar los resultados

- Link a la competencia: https://www.kaggle.com/t/6bc473ee21e840cca2cb3835f3b452c0
- Notebook demo: https://drive.google.com/file/d/1Uihvk3fM0QXJ9flvgTA_0h62xFBSI5_Z/view?usp=sharing