Mecanismos de atención visual en Inteligencia Artificial

Boris Escalante Ramírez

Centro de Estudios en Computación Avanzada, UNAM
&
Laboratorio Avanzado de Procesamiento de Imágenes FI, UNAM

LABORATORIO AVANZADO DE PROCESAMIENTO DE IMÁGENES

Computer vision

David Marr (1976):

"Vision's goal is to produce a useful description of the outside world, free of irrelevant information"

Ian Goodfellow (2016):

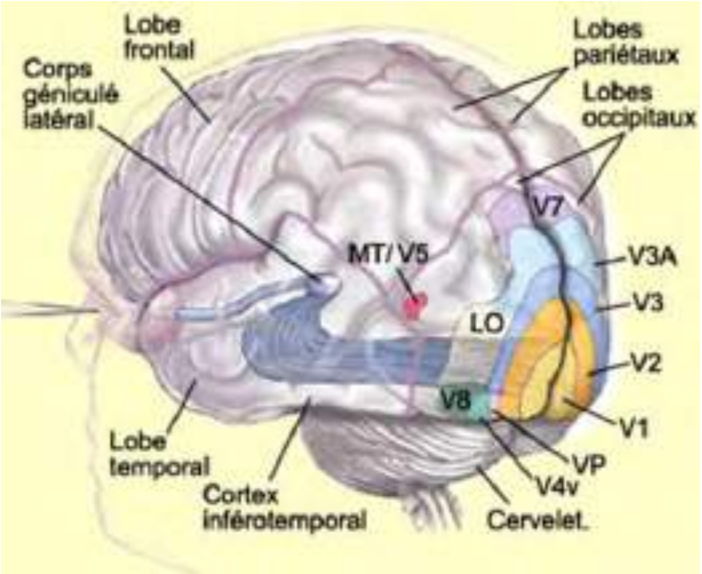
"Computers have long been able to defeat even the best human chess player, but only recently have begun matching some of the abilities of average human beings to recognize objects or speech"

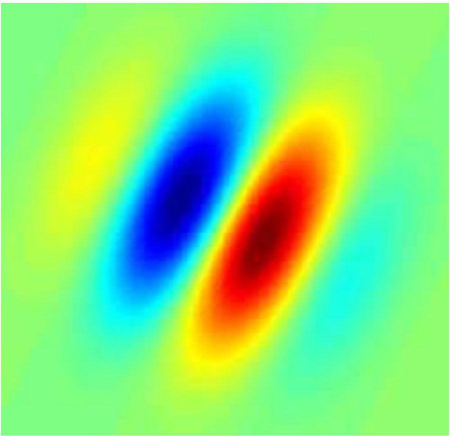
Today's challenge:

"Scene understanding"

Human visual perception

- David Hubel (Nobel 1981): Primary visual cortex (1953) . Simple and complex cells respond to lines, edges, orientation, motion, etc.





Salvador Dalí (1975)

Visual perception

- Witkin (1983), Koenderink (1984). Scale space theory: Multiscale image representation based on Gaussian blurring

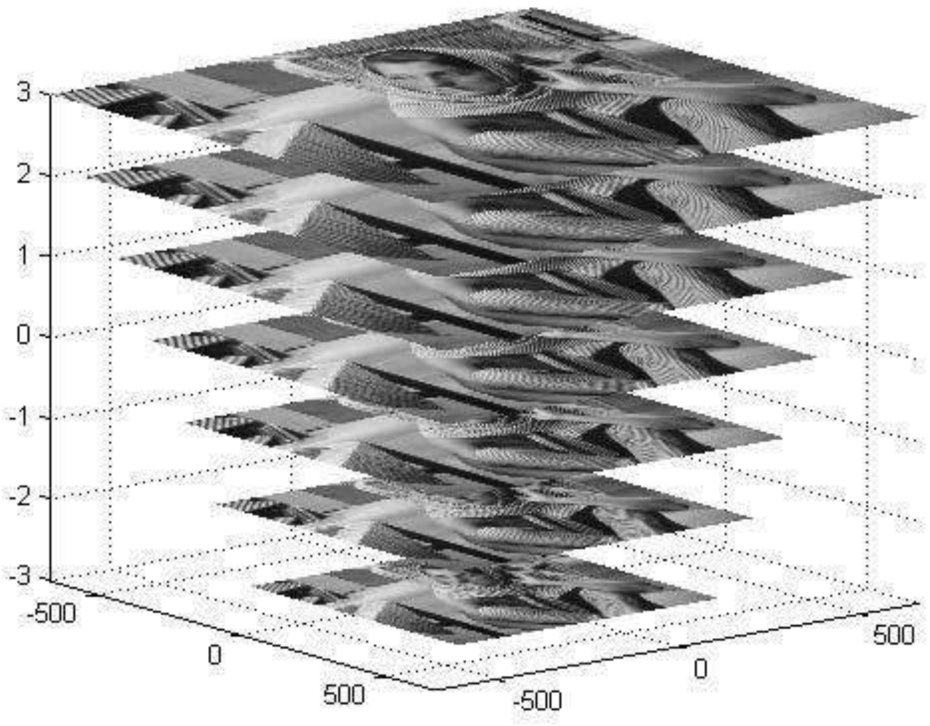


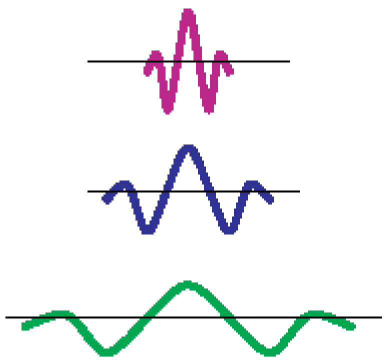
Image representation models

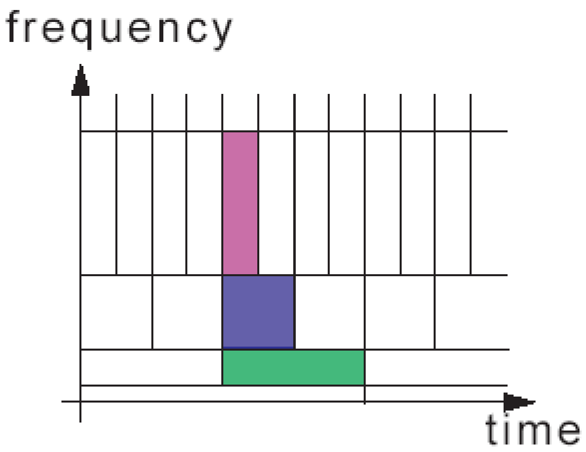
Lineal transformations

- Fourier
- Cosine
- Karhunen-Loève
- **Wavelets**

Wavelets - Advantages

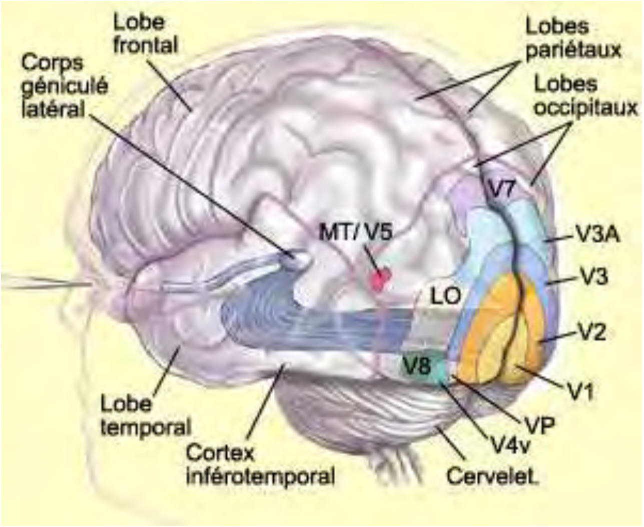
- Good space-frequency localization
- Optimally subsampled
- No redundancy
- Successful in compression



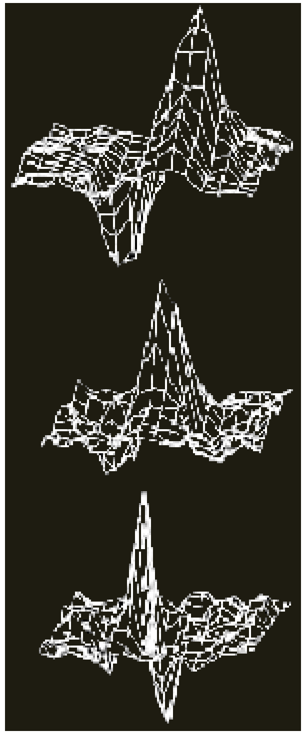


Vision-inspired image transforms

- Gabor
- Cortex
- **Hermite**

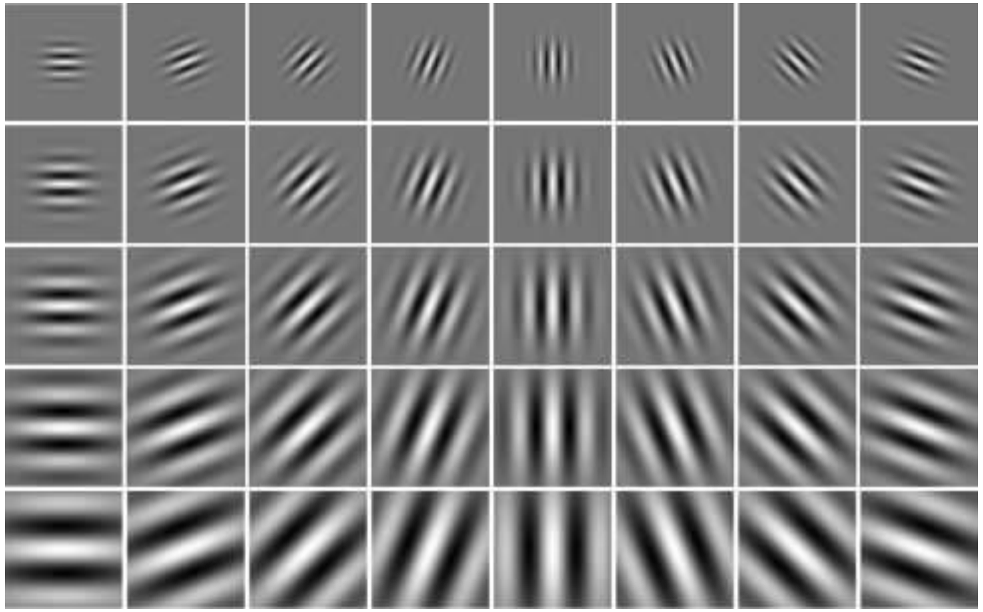


Receptive fields of visual cortex cells



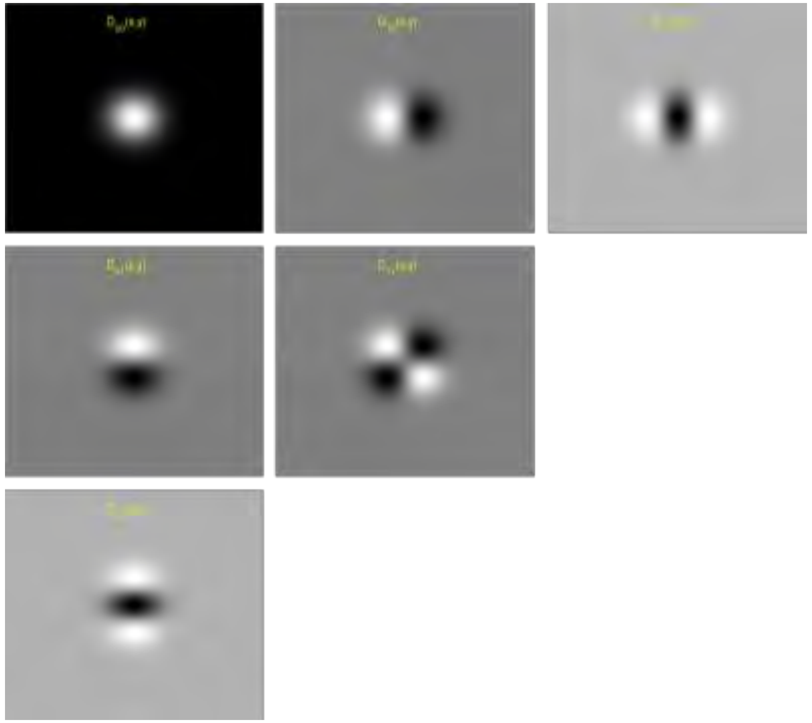
Visual perception

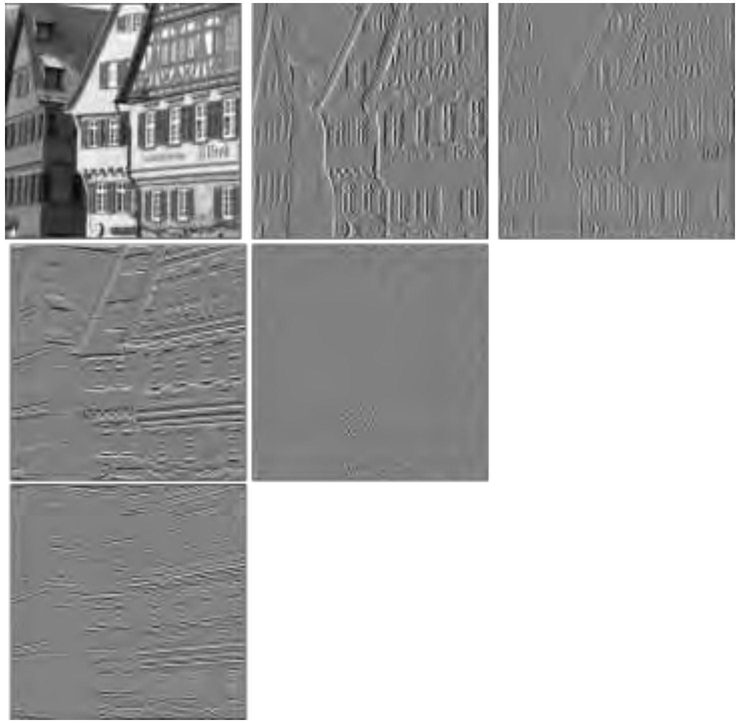
- Marcelja (1980) & Daugman (1985). Gabor functions, good model of simple cells in visual cortex, **but not orthogonal**.



Visual perception

- Martens (1991) Hermite transform. Image representation based on the Gaussian derivative model of human vision





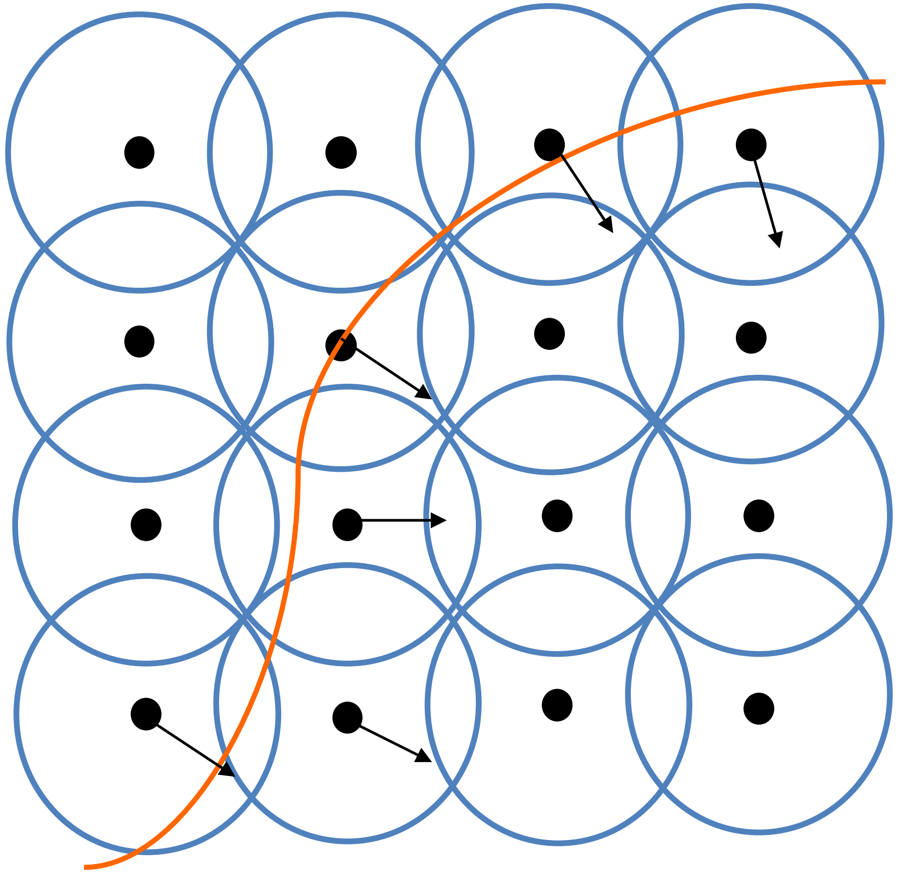
Hermite transform advantages

- Guassian derivatives (human visión model)
- Suited for multiresolution analysis
- Allows rotation (local orientation analysis)
- Redundant (useful for reconstruction)
- Variable subsamping (shift invariant)
- Fast algorithm (NlogN complexity and only additions)

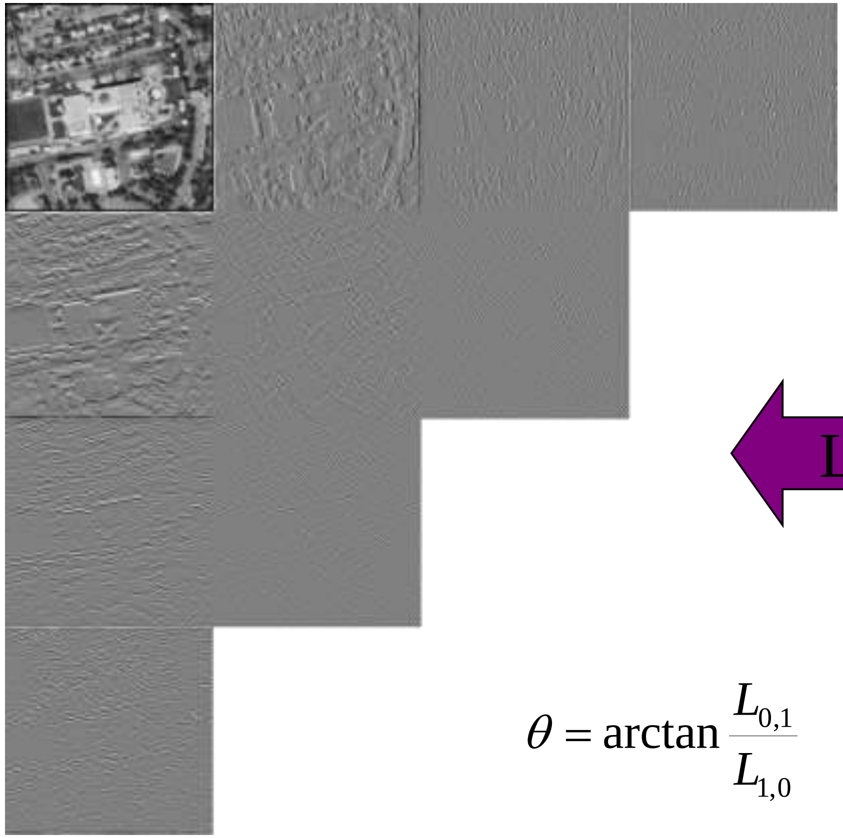
Multi-scale decomposition

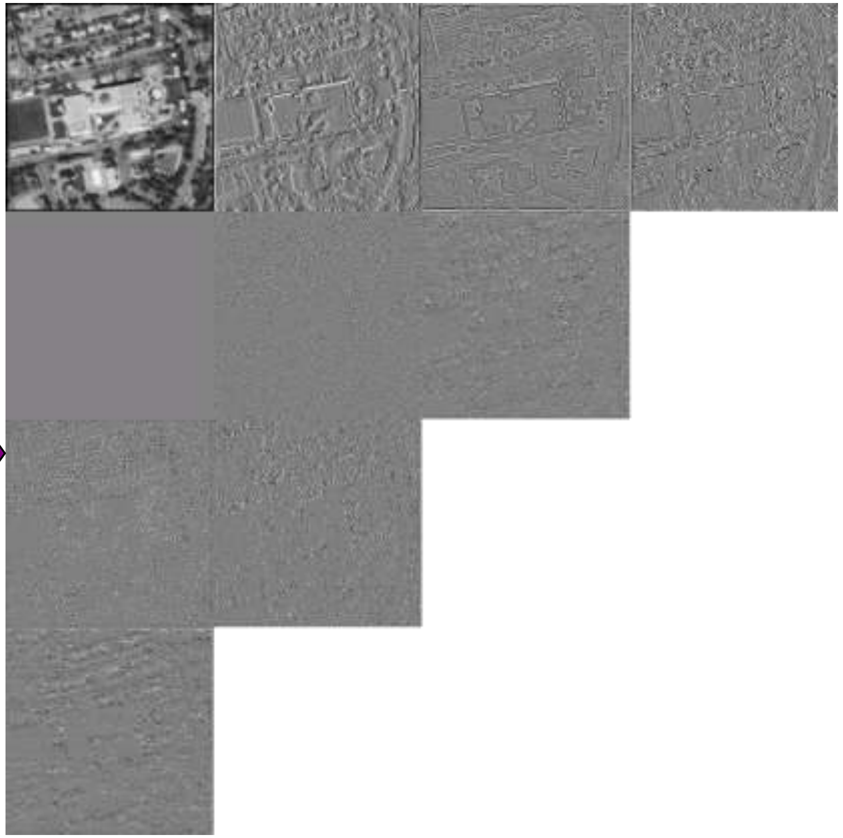
Noise reduction with the HT

Local orientation analysis



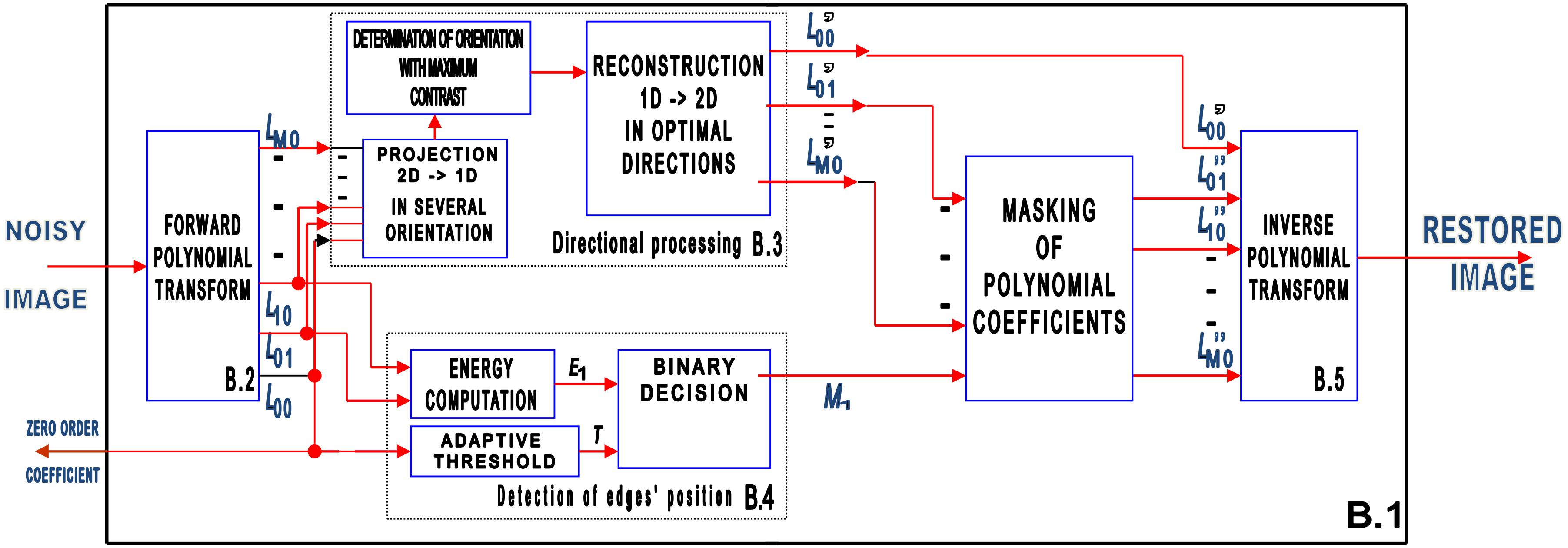
Gradient Rotation





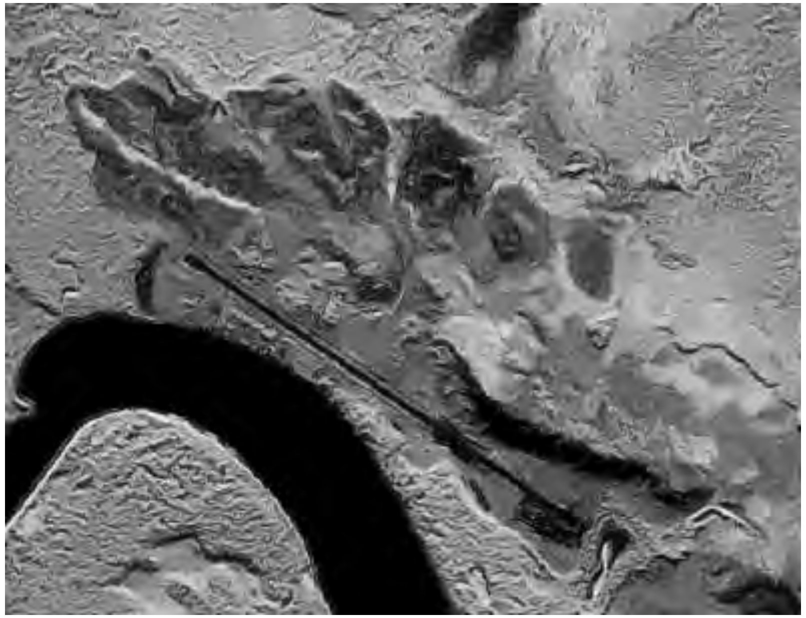
$$\theta = \arctan \frac{L_{0,1}}{L_{1,0}}$$

Noise reduction



Noise reduction

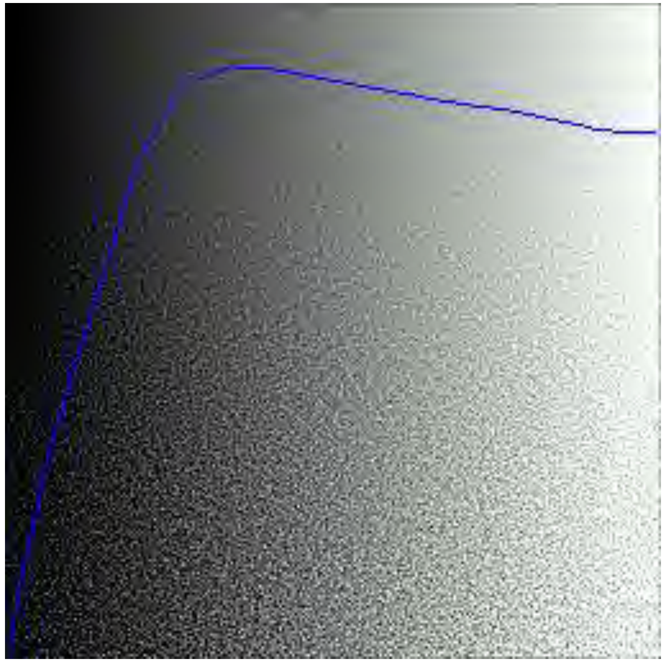
Restored



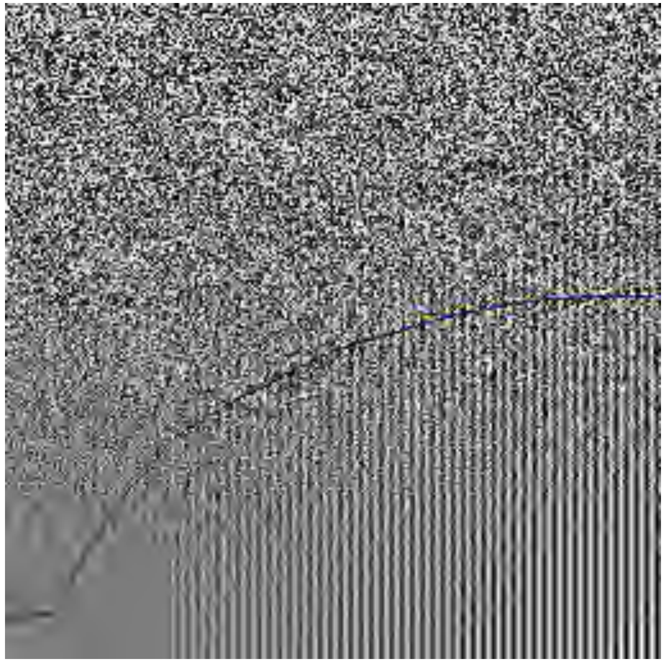
Original

Image Coding

Perceptual visual patterns



Light adaptation threshold



Contrast masking threshold

Visual pattern classification and coding

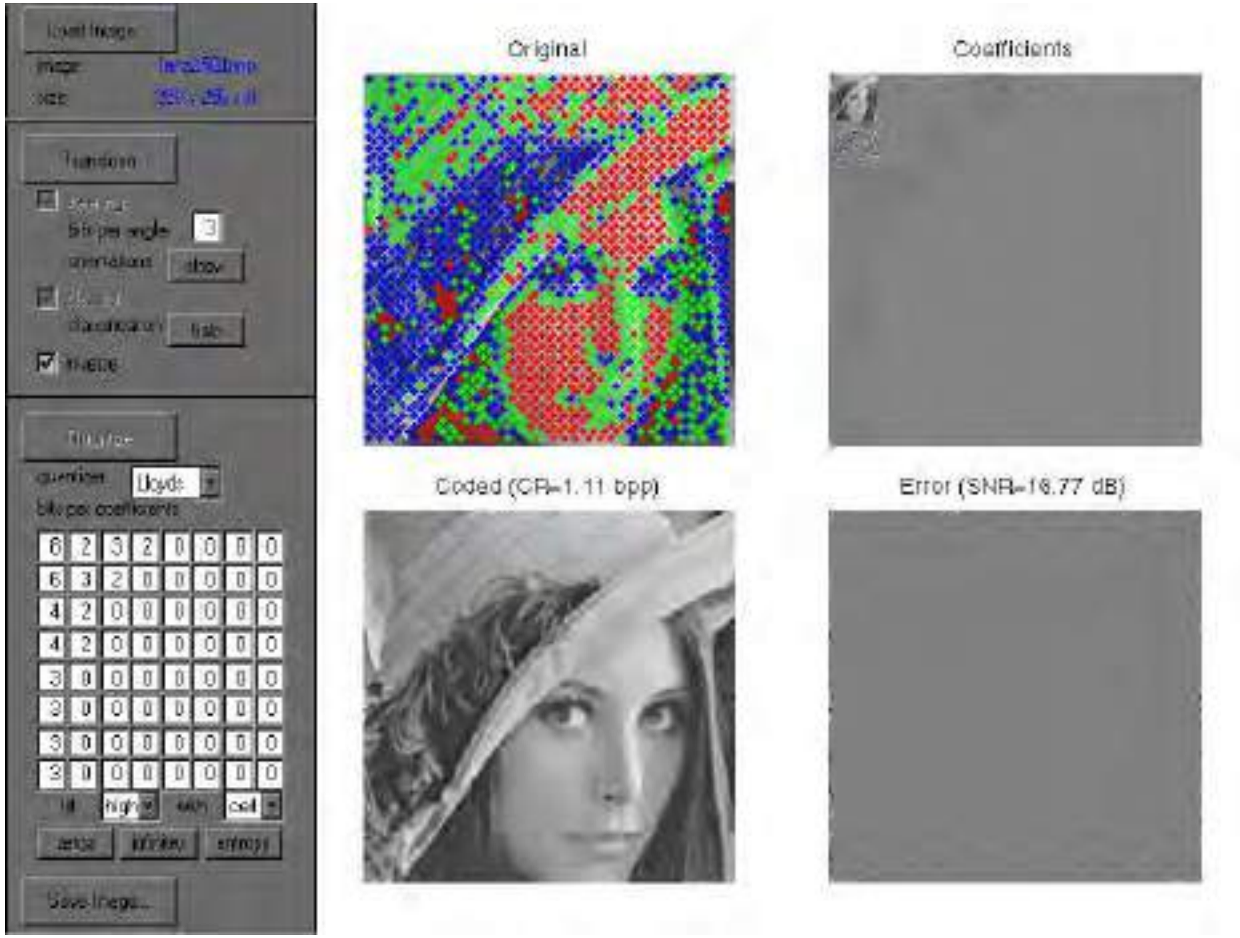
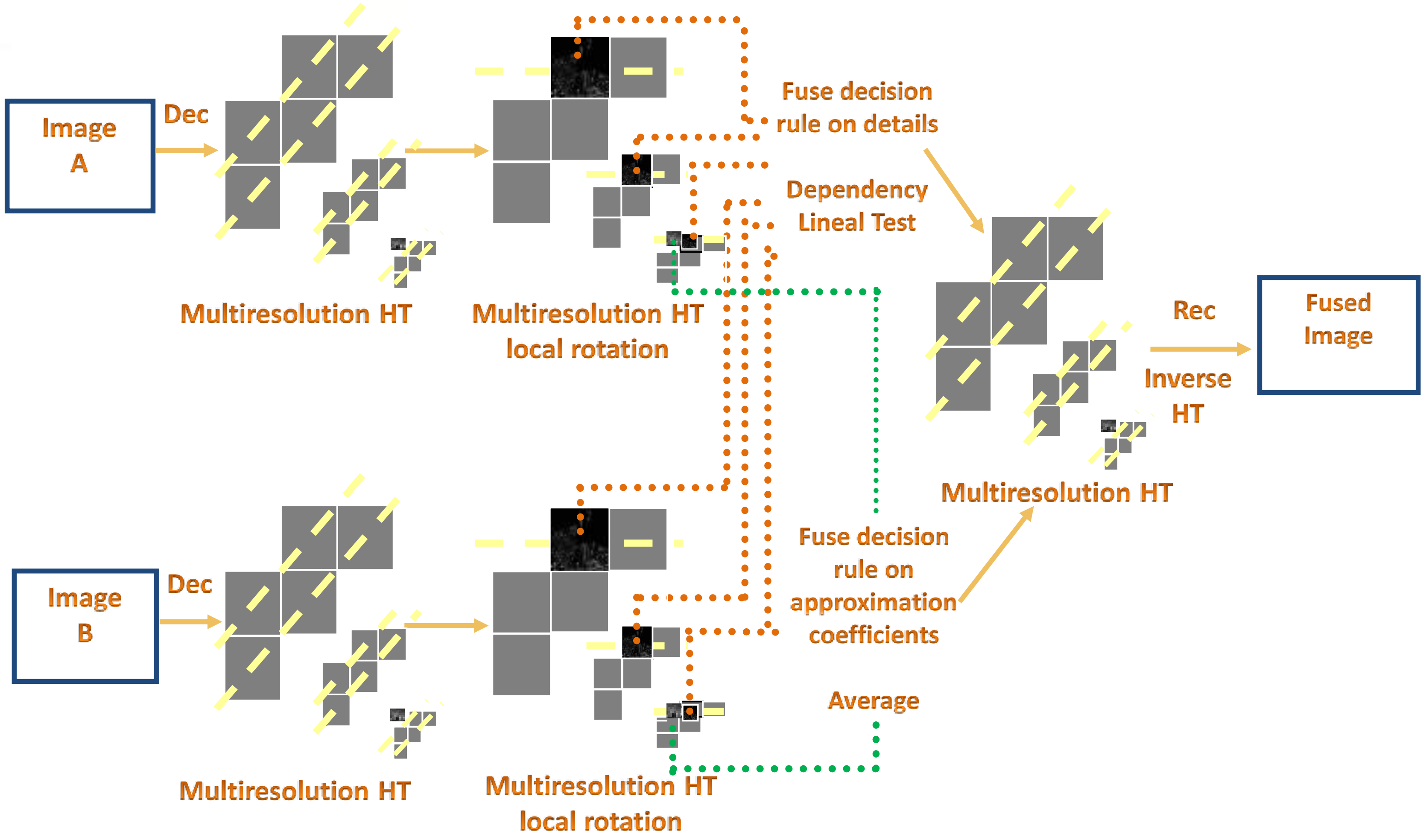
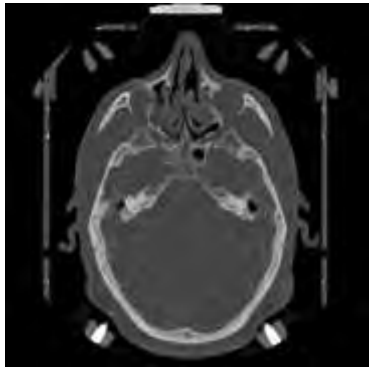


Image fusion

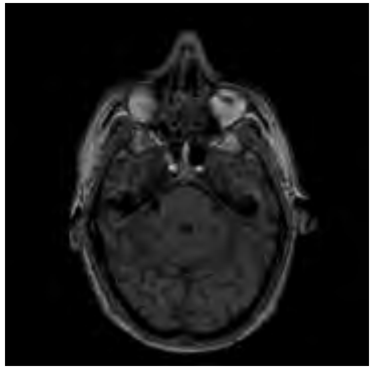
Proposed image fusion algorithm



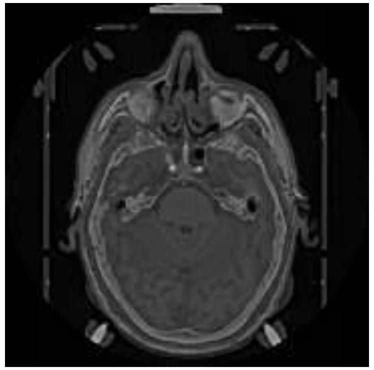
Results of image fusion in multimodal images



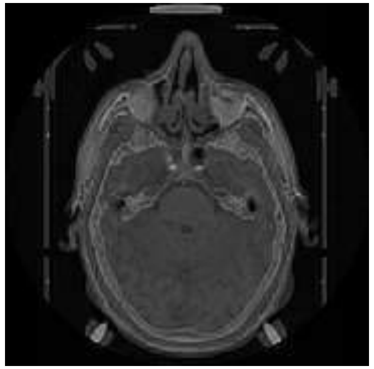
(a)



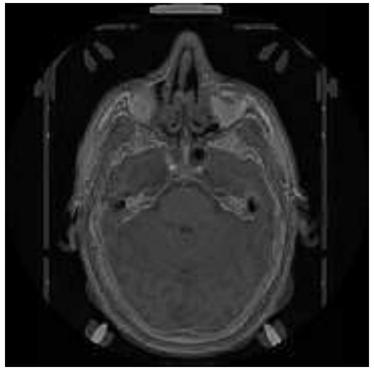
(b)



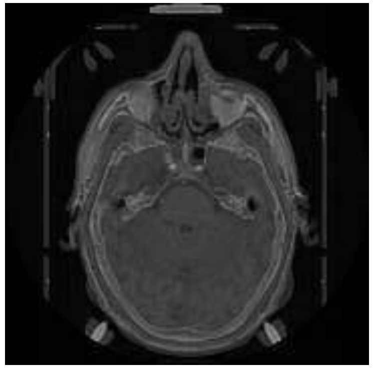
(c)



(d)



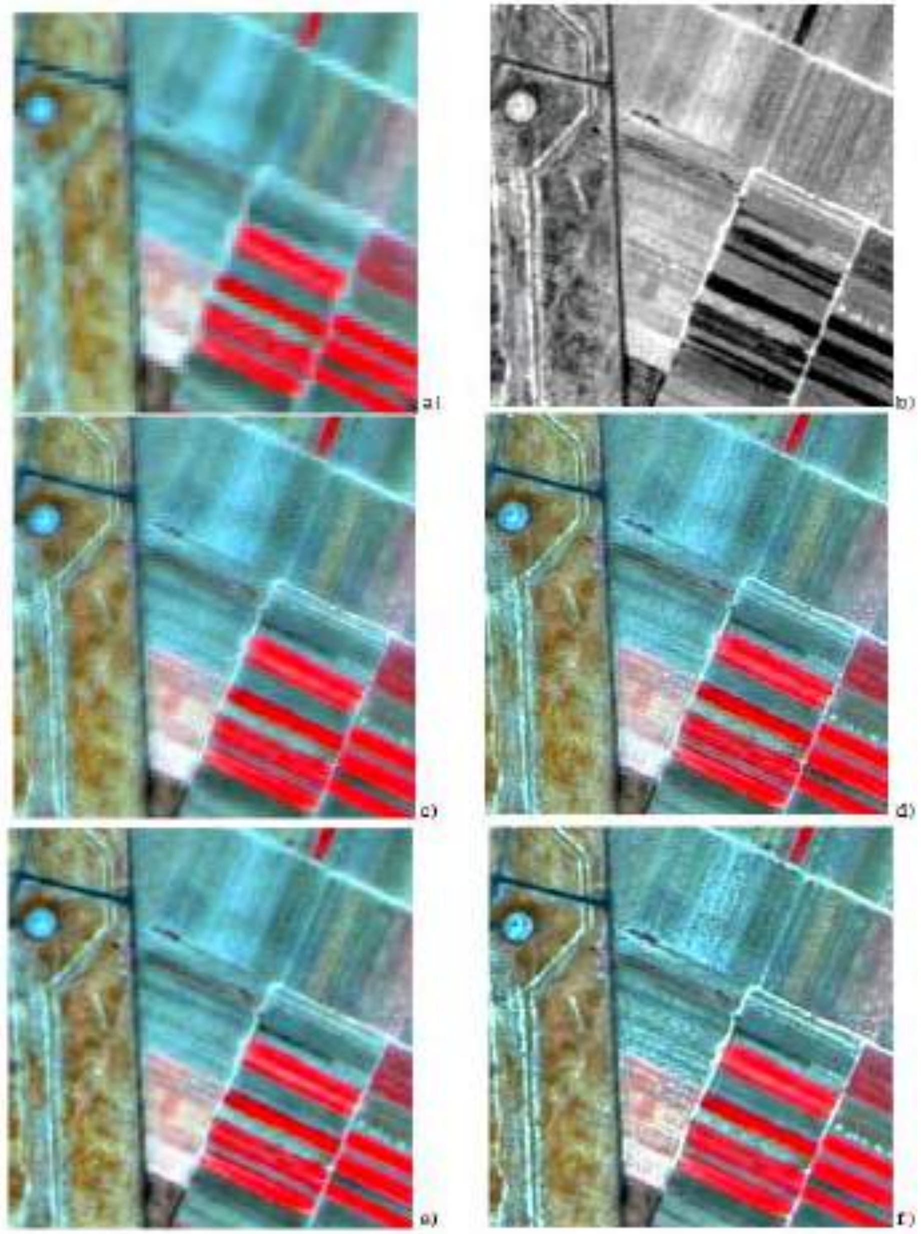
(e)



(f)

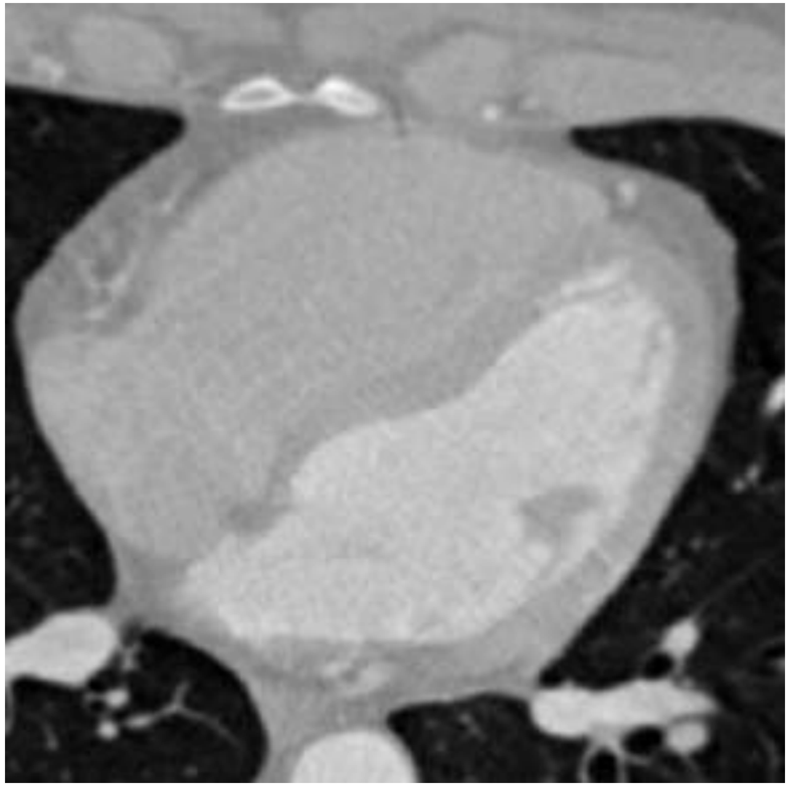
(a) CT and (b) MR are the source images, fused images using c) TH, d) TW, e) CW and f) CUW.

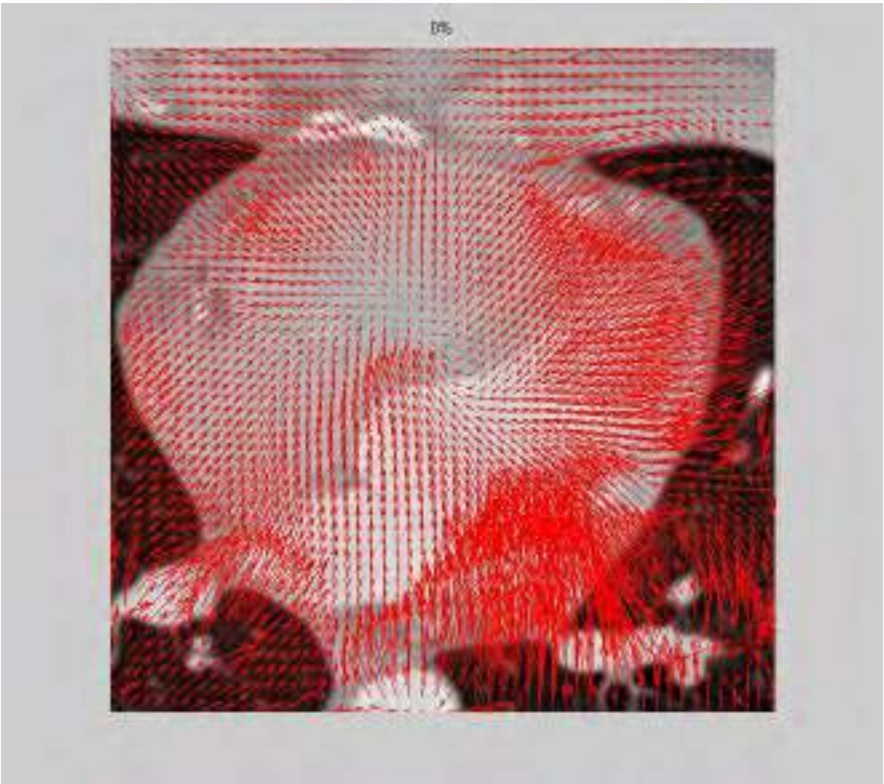
Landsat multiespectral & panchromatic fusion: Pansharpening



Optic flow estimation

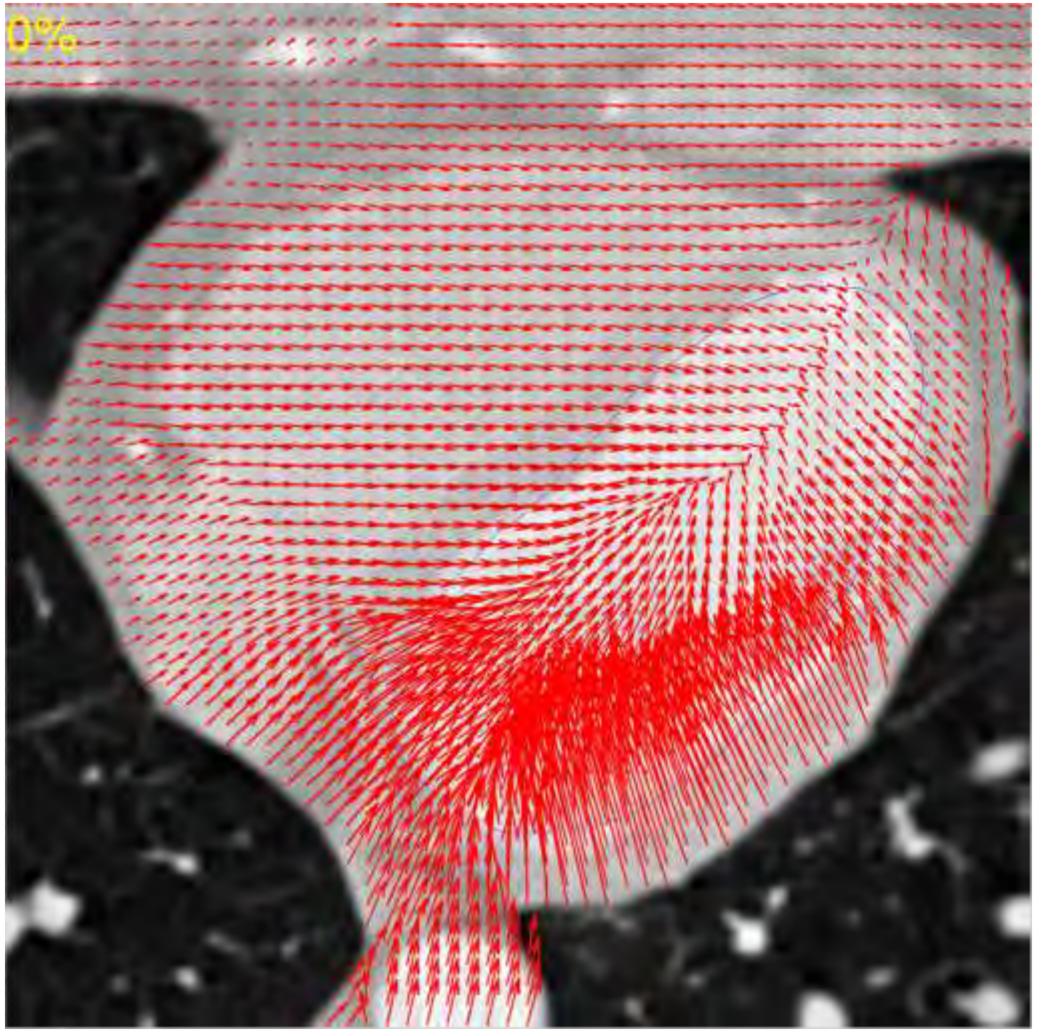
CT-heart flow estimation



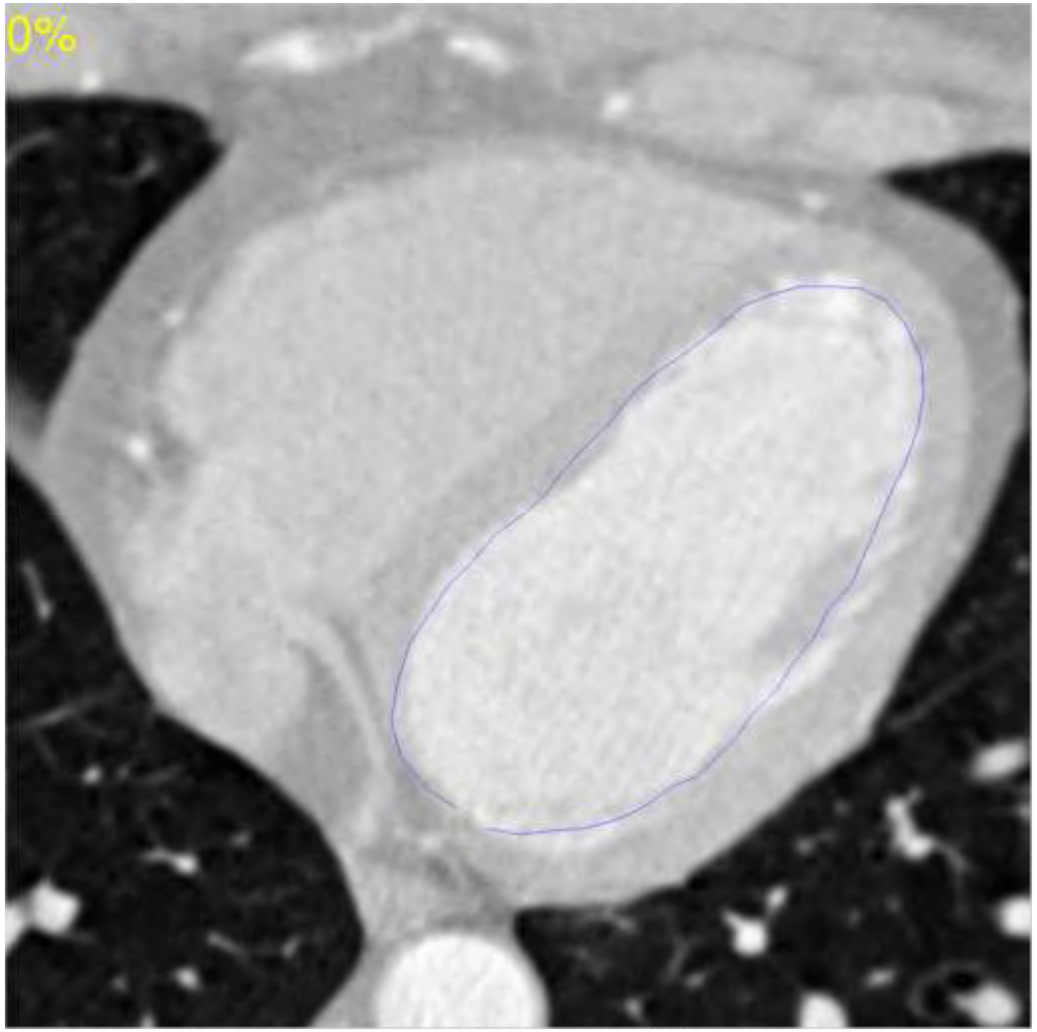


Computed tomography heart sequence

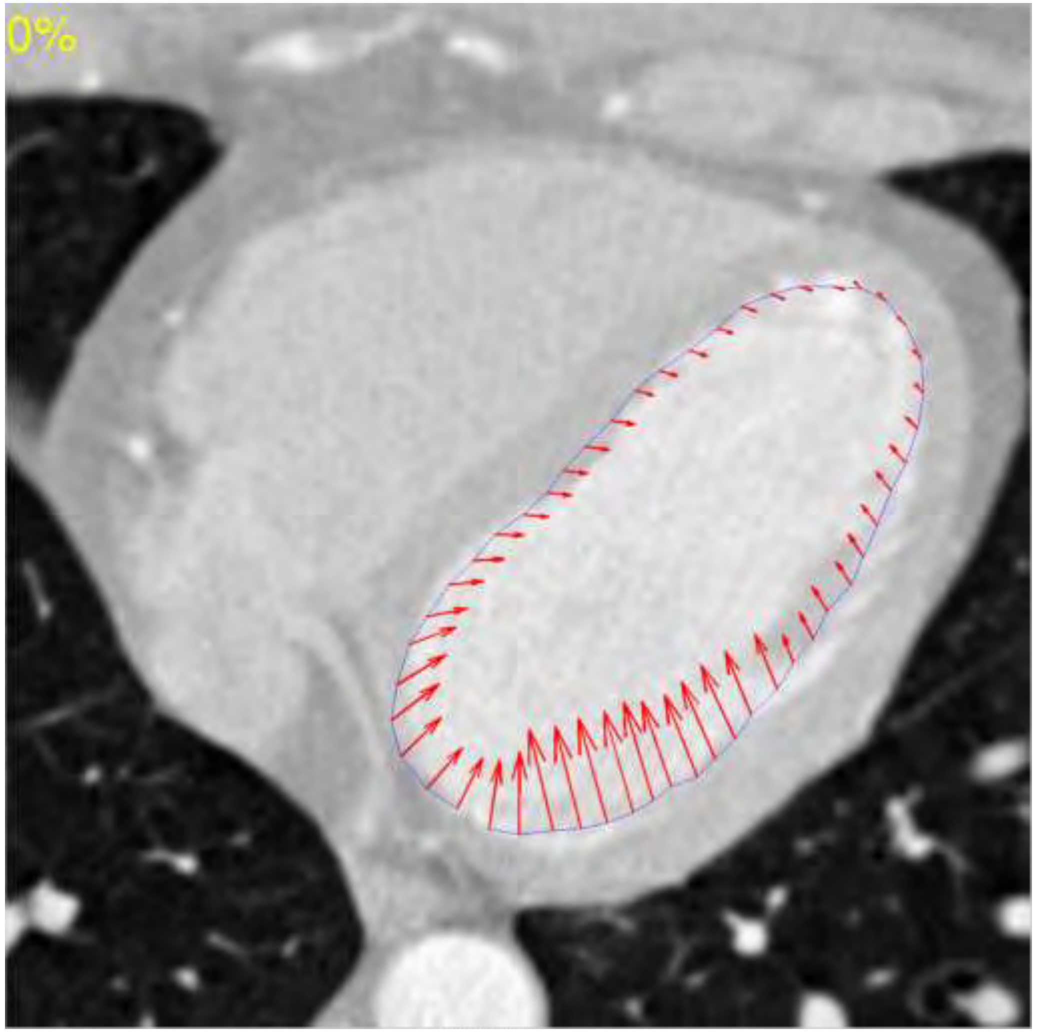
CT-heart flow estimation



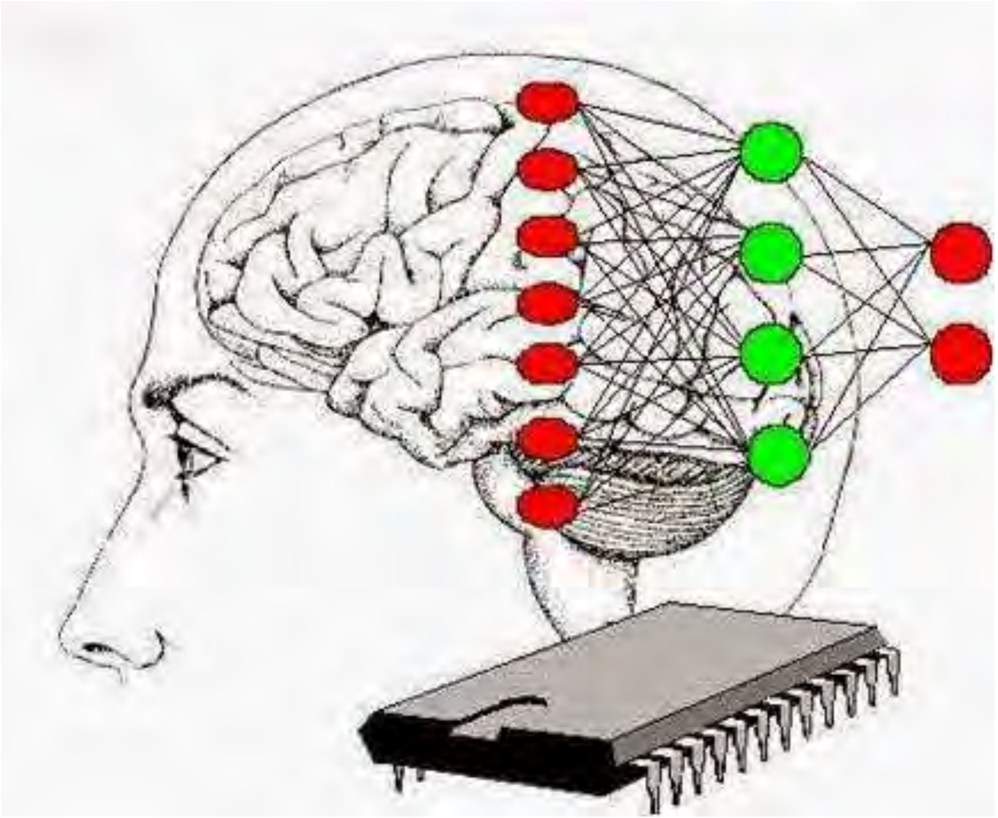
Left ventricle segmentation

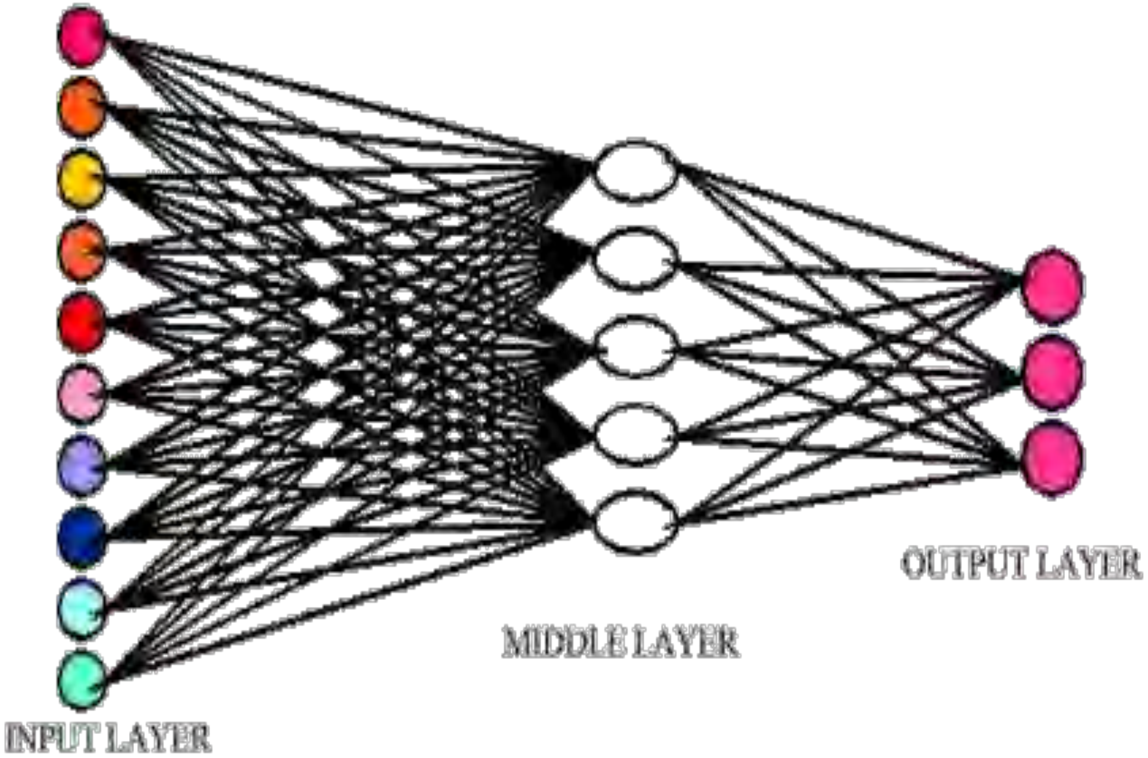


Left ventricle motion estimation



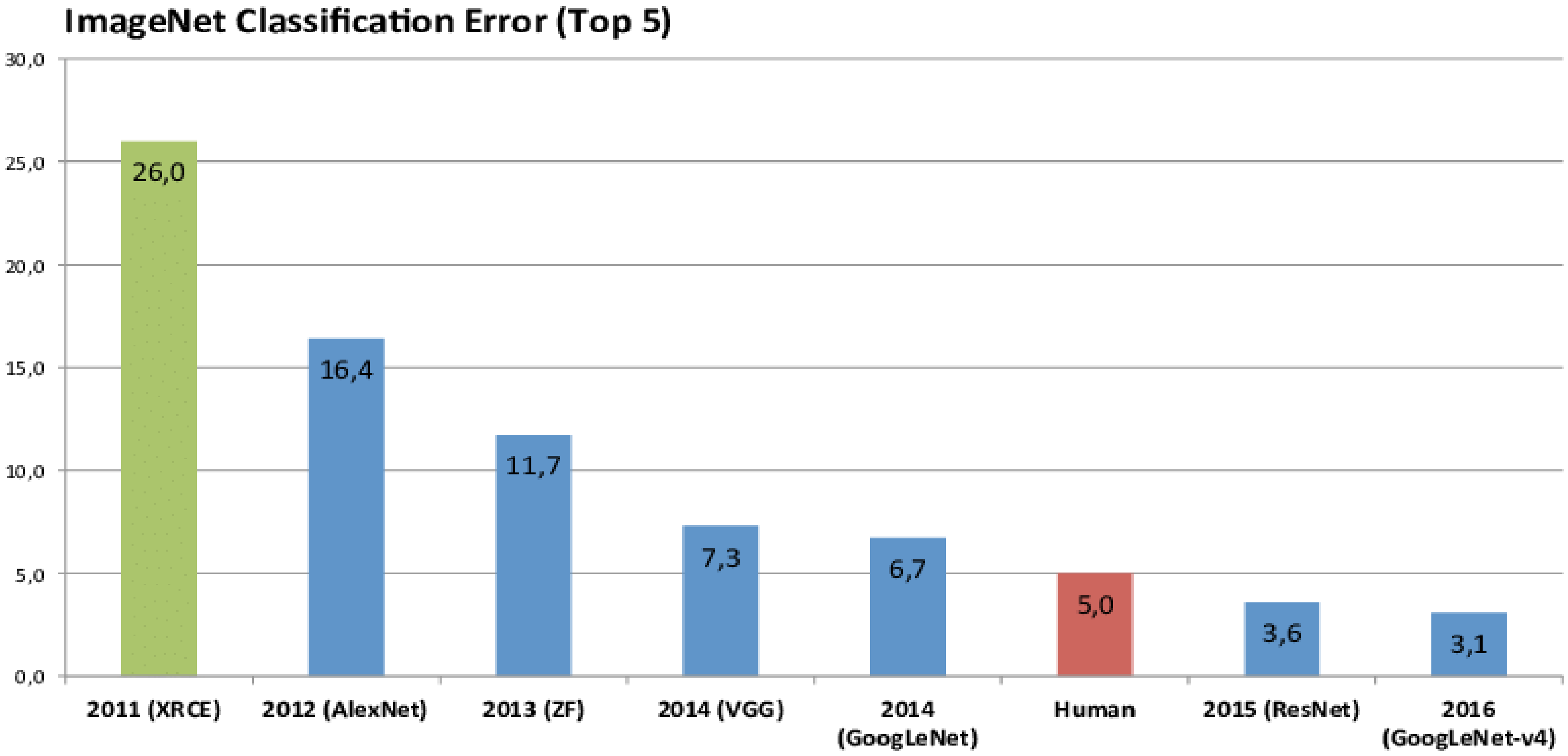
Artificial neural networks inspired in human neural networks



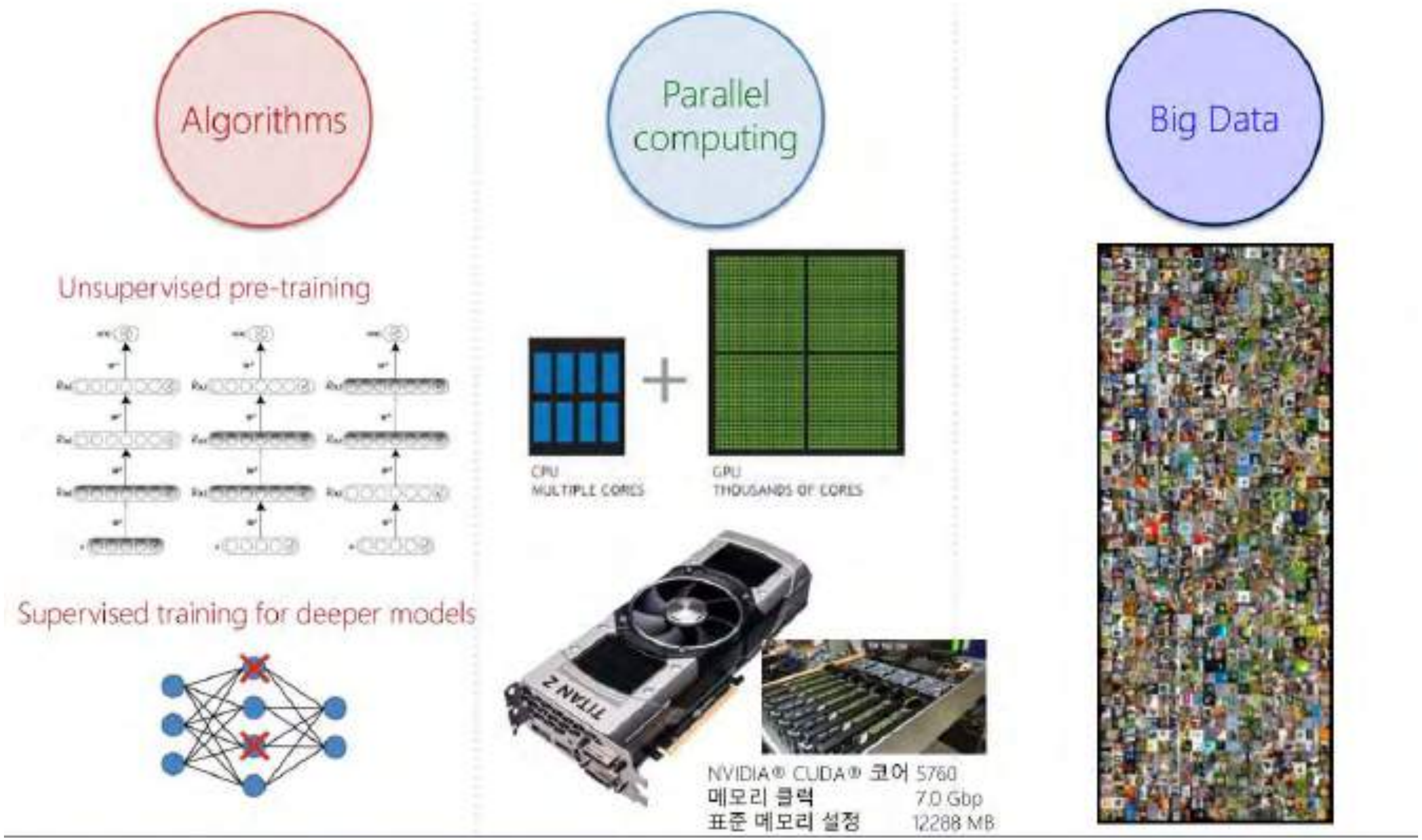


Deep learning evolution

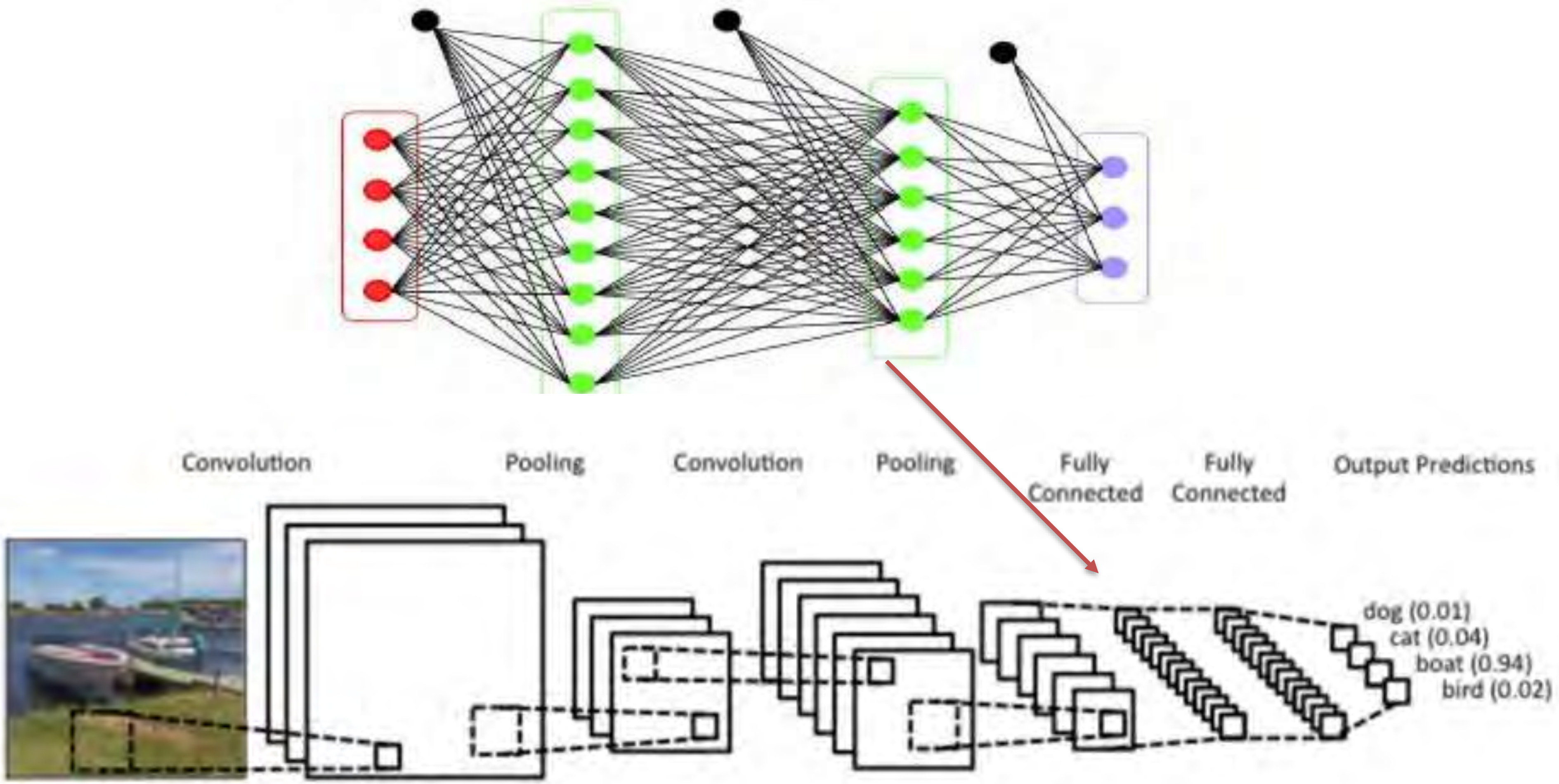
1,000 classes, 14,197,122 images

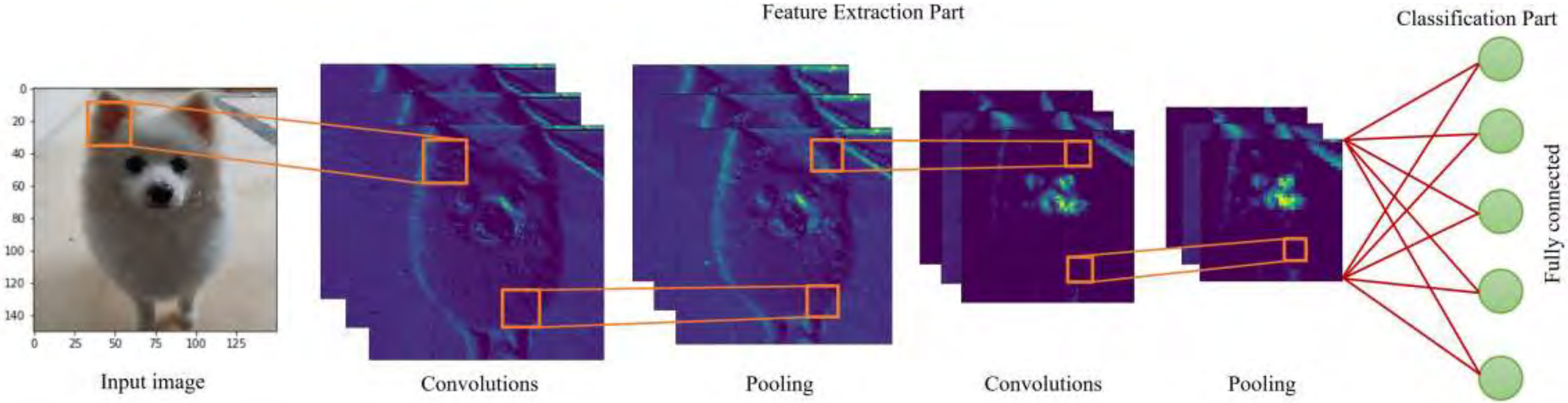


Why are they good?



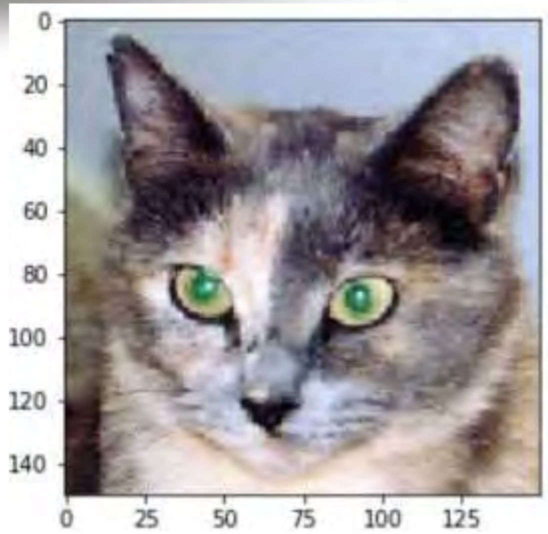
Convolutional neural networks (CNN): visual perception !!!!

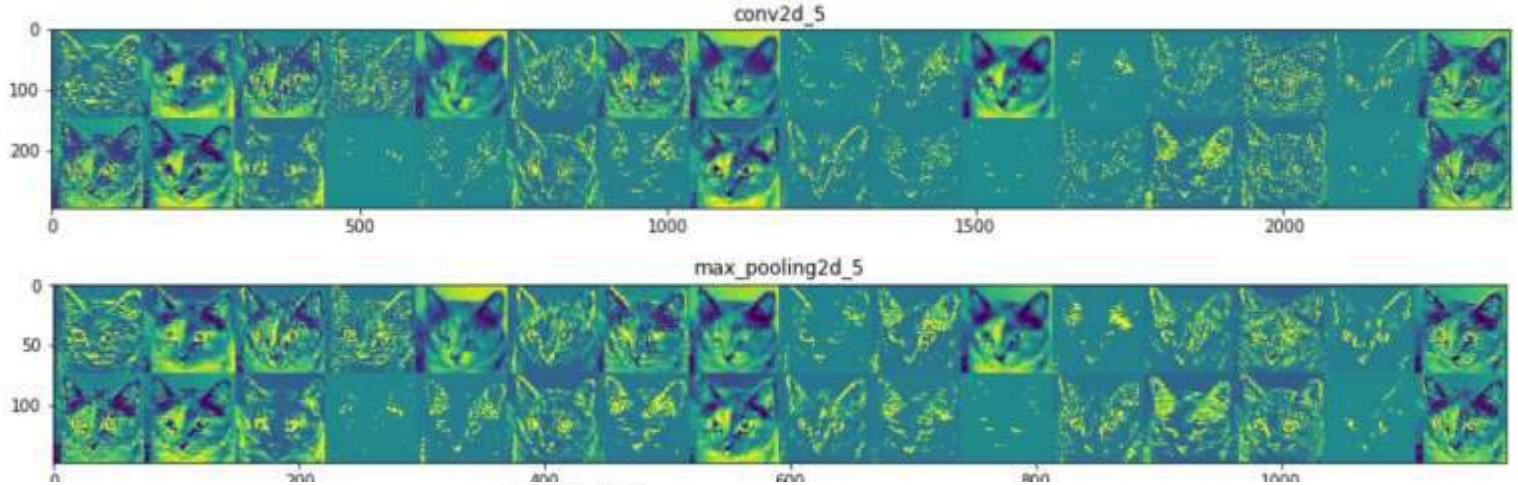


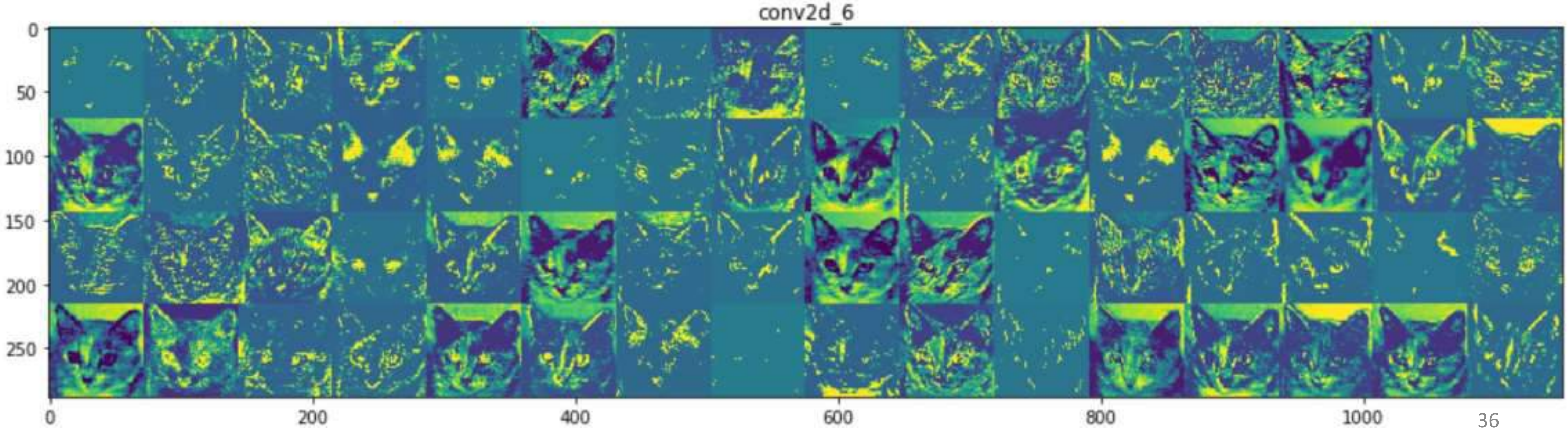


Feature extraction (convolutional NN) followed by a classifier (fully connected NN).

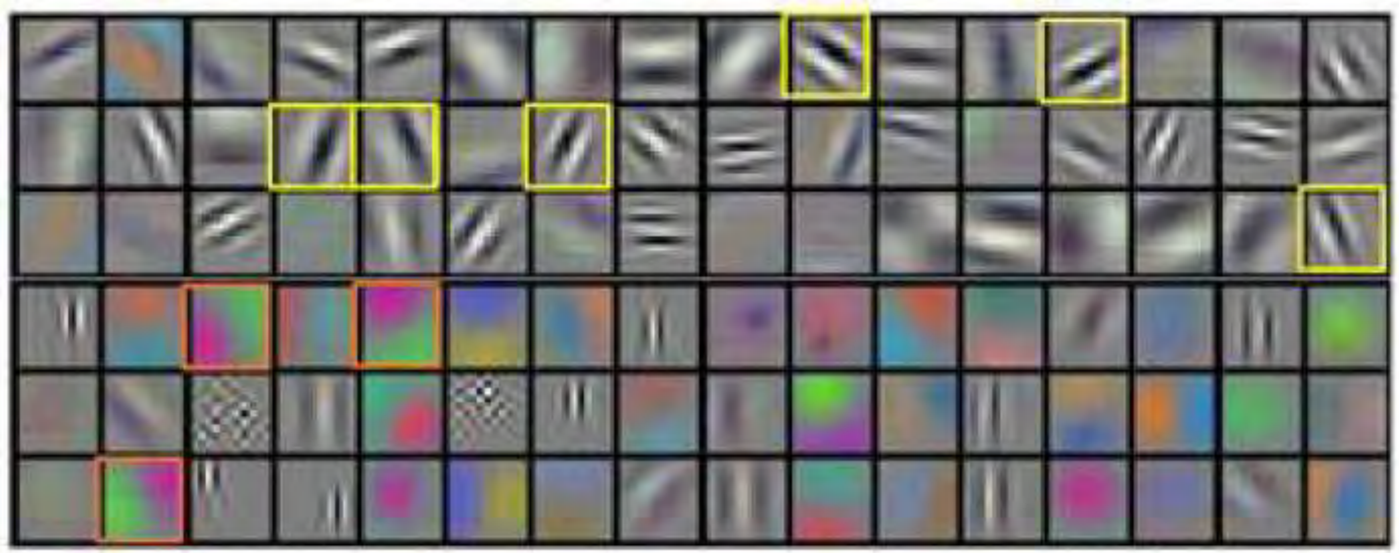
Feature maps



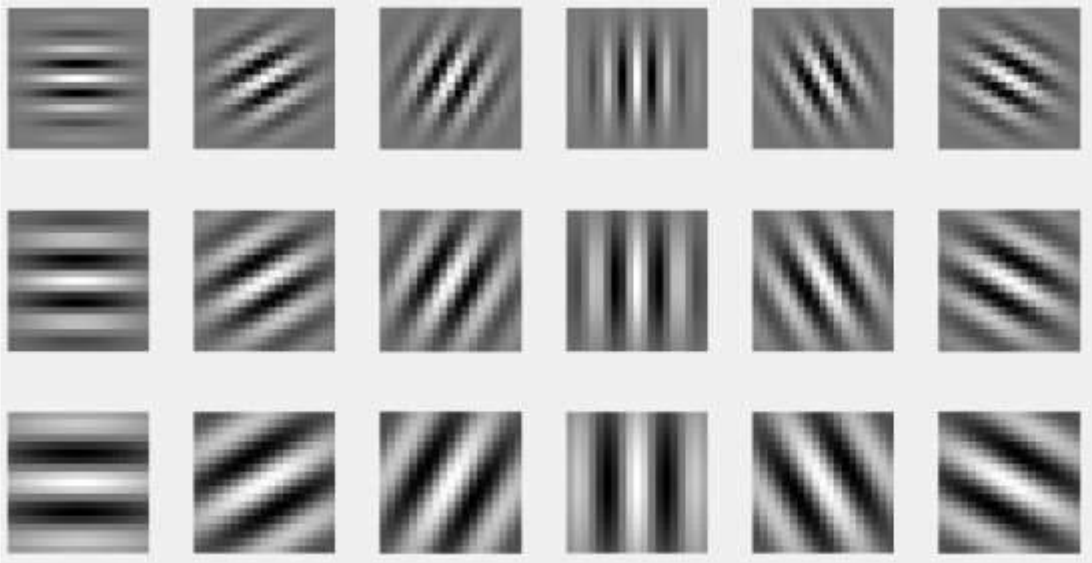




LaPI
LABORATORIO AVANZADO DE PROCESAMIENTO DE IMÁGENES

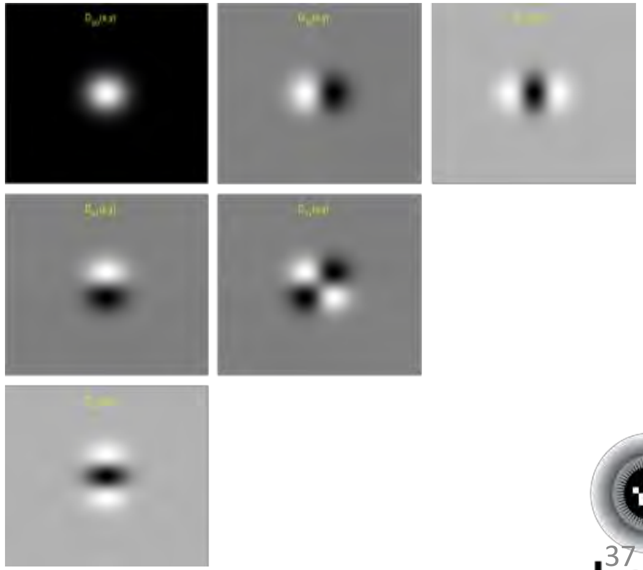


FAlexNet filters in hidden layers



Gabor filters

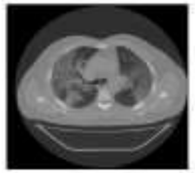
Hermite filters



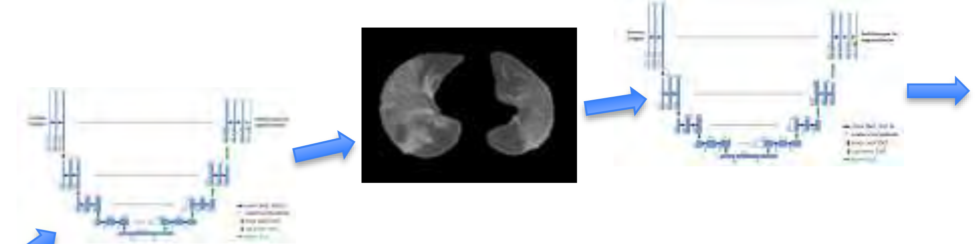
COVID-19 diagnosis aid tool

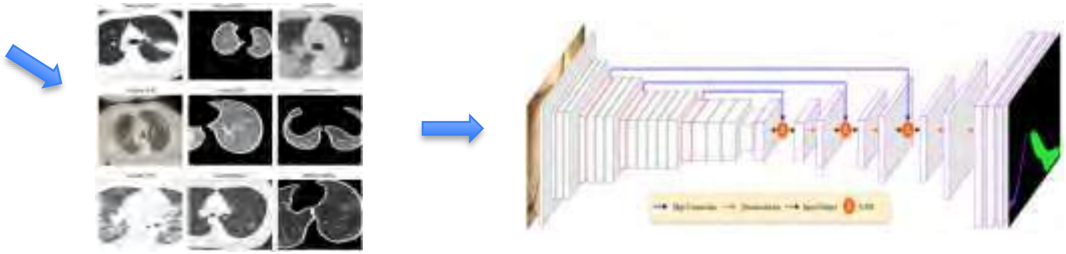
- Automatically analyzes chest CT or X-ray images
- Uses computer vision and artificial intelligence techniques
- Accuracy of 98%
- Outcome 1: Probability of COVID-19
- Outcome 2: Detection and localization of characteristic lesions (ground glass and consolidations) (CT only)
- First free open website in the world for COVID-19 detection
- First algorithm for detection of COVID-19 by CT and Rx made in Mexico
- www.imagensalud.unam.mx

Computer assisted diagnosis of COVID-19



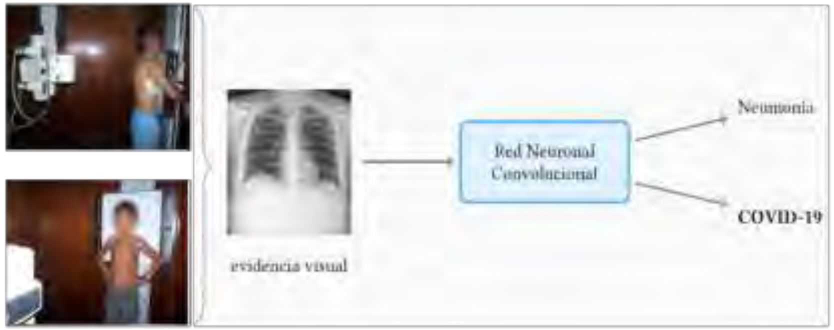
DICOM
JPG
TIF
PNG



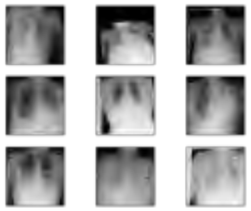


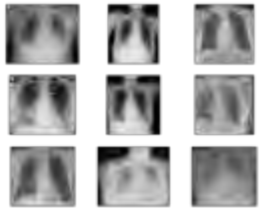
PA

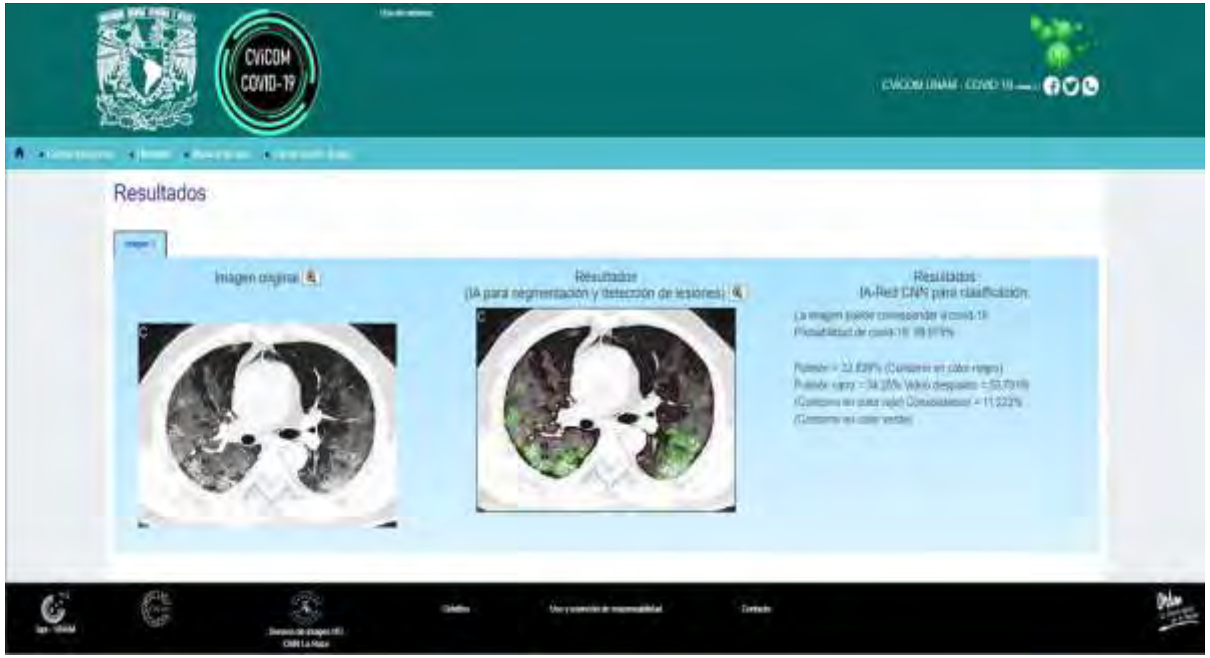
AP



AUCROC	Exactitud	Especificidad	Sensibilidad	Precisión	F1
98.16	98.50	99.50	78.33	88.68	83.19

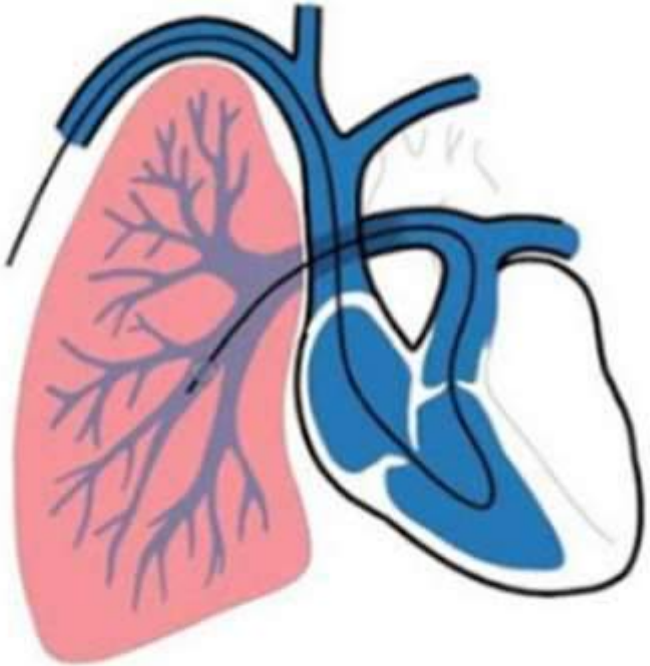


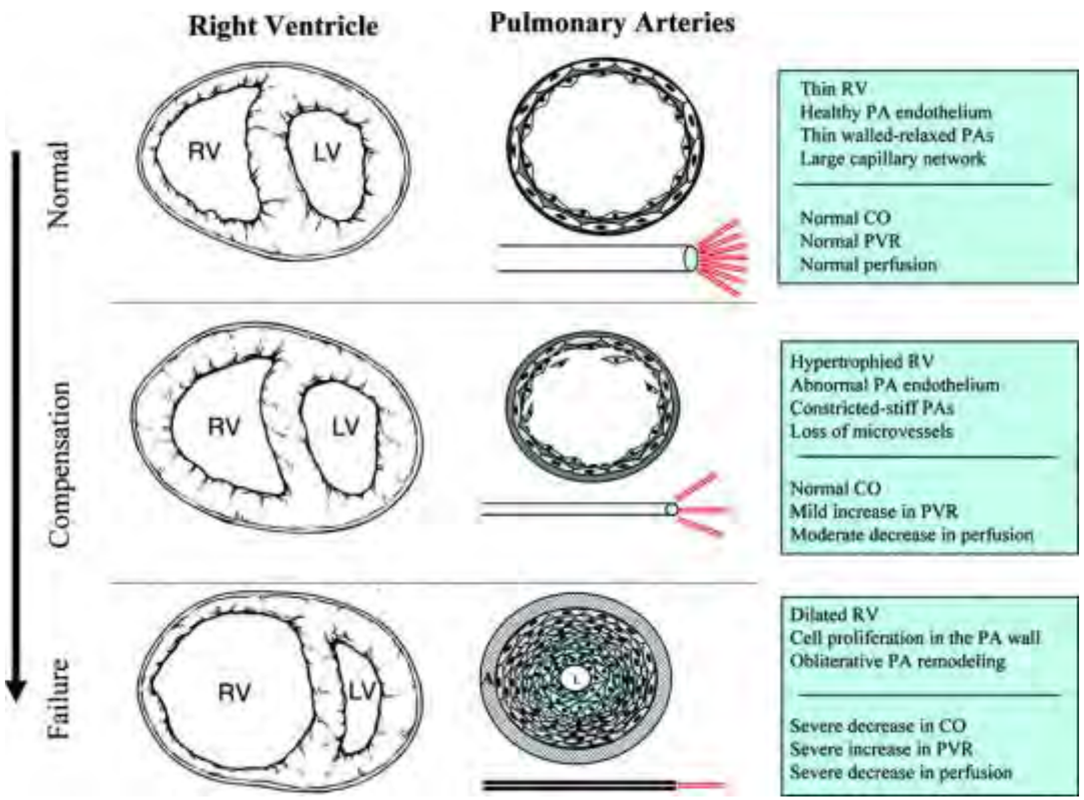


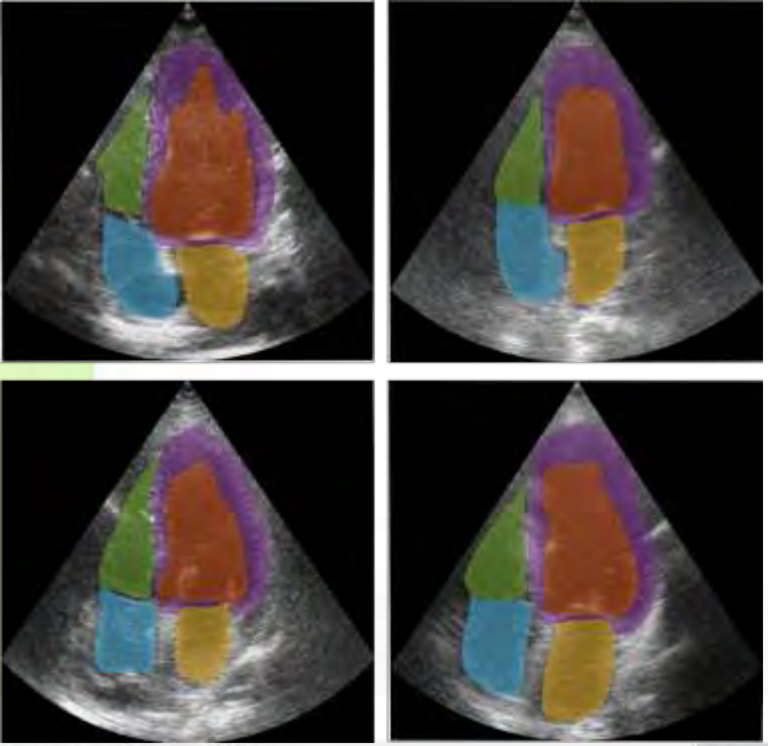


INTELLIGENT SYSTEM FOR ASSISTED MEDICAL DIAGNOSIS FOR CARDIOVASCULAR DISEASES

- Centro Médico Nacional "20 de Noviembre" ISSSTE (Dra. Rocío Aceves).
- Gold standard: pulmonary artery catheterization via right heart chambers
- It is risky, costly, and requires high specialization

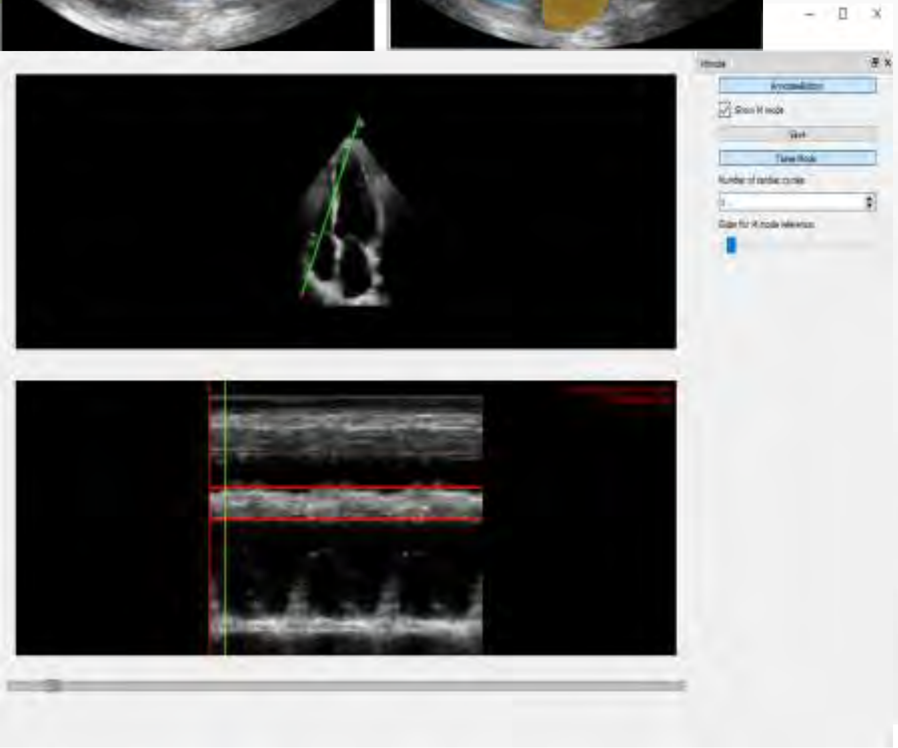


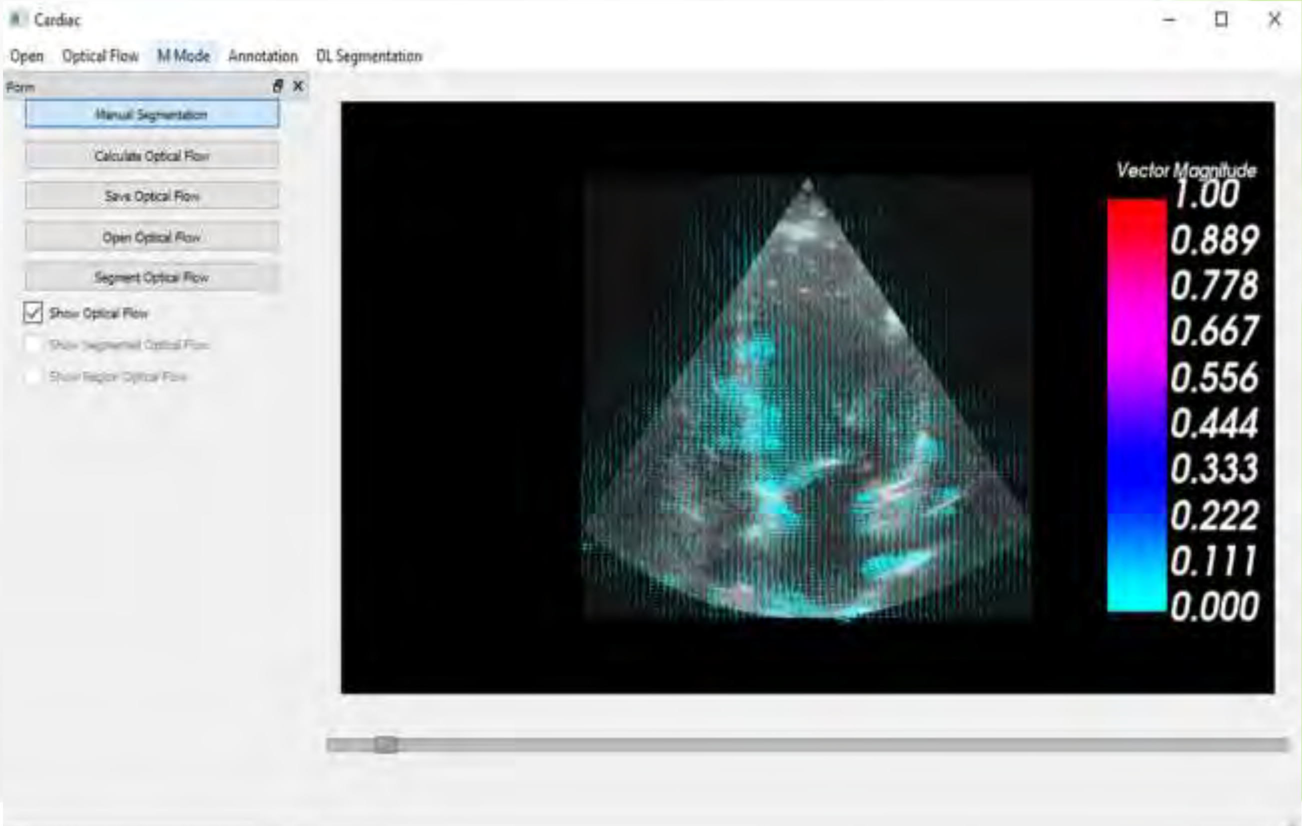




Intelligent system for assisted medical diagnosis of cardiovascular diseases", SYSCARDIO

Automatically analyzes echocardiography images using computer vision and artificial intelligence.





Transformers

Un antecedente:

Neural Machine Translation by Jointly Learning to Align and Translate

2014, Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio

Neural Machine Translation by Jointly Learning to Align and Translate

2014, Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio

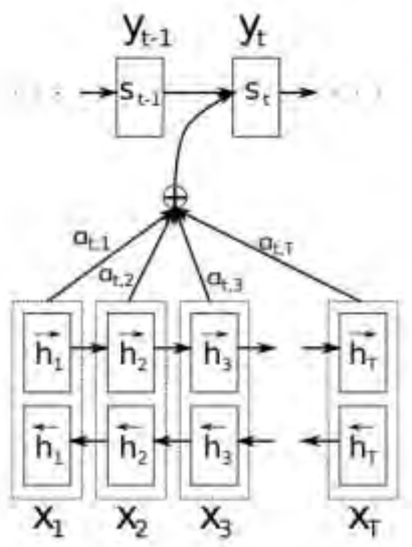
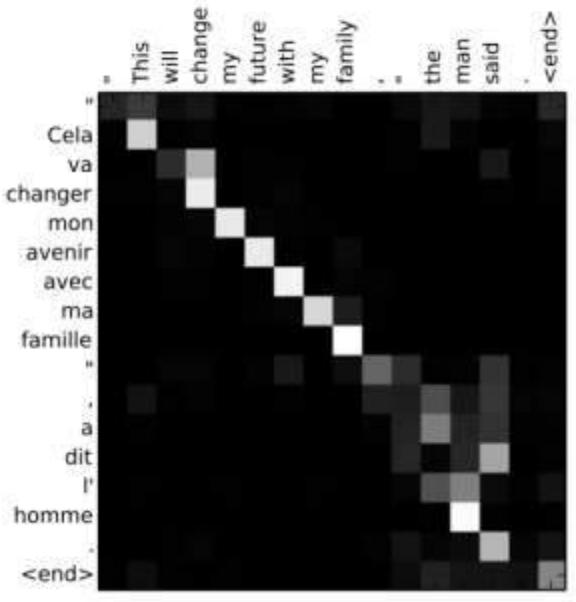


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.



(d)

The probability α_{ij}, or its associated energy e_{ij}, reflects the importance of the annotation h_j with respect to the previous hidden state s_{i-1} in deciding the next state s_i and generating y_i. Intuitively, this implements a mechanism of attention in the decoder. The decoder decides parts of the source sentence to pay attention to. By letting the decoder have an attention mechanism, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed-length vector. With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly.

La corazón del transformer es la ecuación de “atención”:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Attention Is All You Need - The Transformer architecture

2017, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

La entrada debe ser pasar por un proceso de tokenization y creación de un embedding

El texto de entrada se divide primero en partes. Pueden ser caracteres,

palabras o “tokens”:

`El perro está dormido" -> [El_] [perro_] [esta] [dorm] [ido_]`

Los tokens se mapean a un "vocabulario", a esto se le llama “numericalizar”:

`[El_] [perro_] [esta] [dorm] [ido_] -> [3 721 68 1337 42]`

Cada entrada del vocabulario corresponde a un vector de dimensión “d” definido por el modelo (embedding)

`[3 721 68 1337 42] -> [ [0.123, -5.234, ...], [...], [...], [...], [...] ]`

Se debe codificar la posición de cada token (positional encoding)

El idioma escrito no es invariante a la permutación.

Se necesita codificar la posición de cada palabra; se suma un embedding de posición

`[El_] + 10   [perro_] + 20  [está] + 30`

1. Los vectores, k and v se derivan de la entrada de embeddings x: $k = W_k \cdot x$ $v = W_v \cdot x$

El vector de query q se puede obtener de la entrada de embeddings x. En tal caso se llama "self attention": $q = W_q \cdot x$

También se puede obtener de una segunda entrada y. Entonces se llama “cross-attention”: $q = W_q \cdot y$

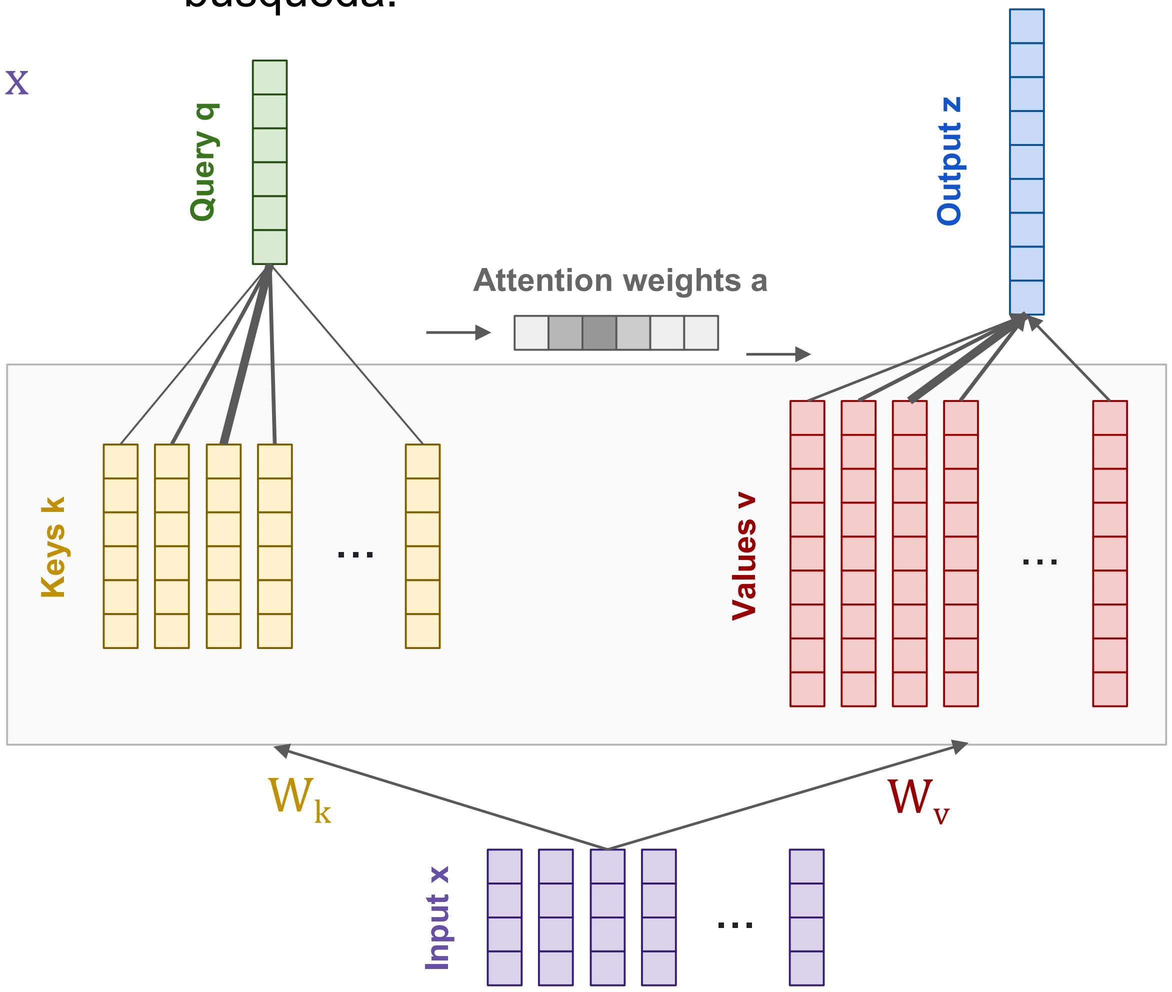
2. Se le llama pesos de atención (attention weights, $a_{1:N}$)

$\hat{a}_i = q \cdot k_i$

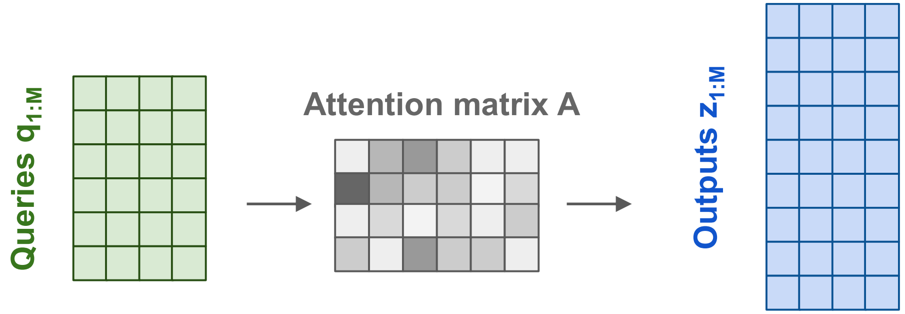
Se normalizan por una función softmax: $a_i = e^{\hat{a}i} / \sum_j e^{\hat{a}j}$

3. La salida z se obtiene ponderando el vector de pesos de atención con el vector de values $v_{1:N}$: $z = \sum_i \hat{a}_i v_i = \hat{a} \cdot v$

Qué tan parecida es nuestra query a cada key. La idea es extraer en la salida el vector de value correspondiente al vector key que es semejante a nuestra búsqueda.



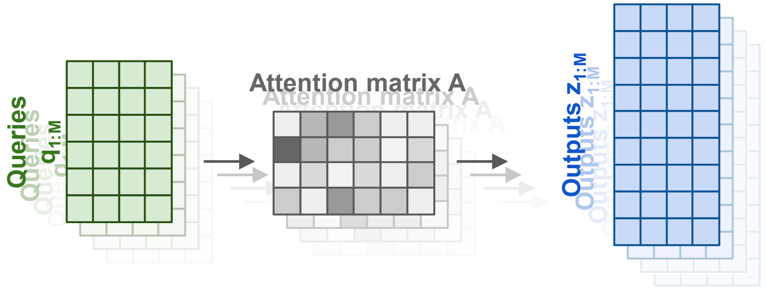
1. Usualmente se utilizan muchos vectores de queries $q_{1:M}$, no sólo uno. Al agruparlos se forma la matriz de atención (attention matrix $A_{1:N,1:M}$) lo que produce muchas salidas:



$$z_{1:M} = \mathrm{Attn}(q_{1:M}, x) = [\mathrm{Attn}(q_1, x) \mid \mathrm{Attn}(q_2, x) \mid \ldots \mid \mathrm{Attn}(q_M, x)]$$

2. Usualmente se usa algo llamado "multi-head" attention. Esto quiere decir que la operación se realiza K veces y los resultados se concatenan a lo largo de la dimensión de características.

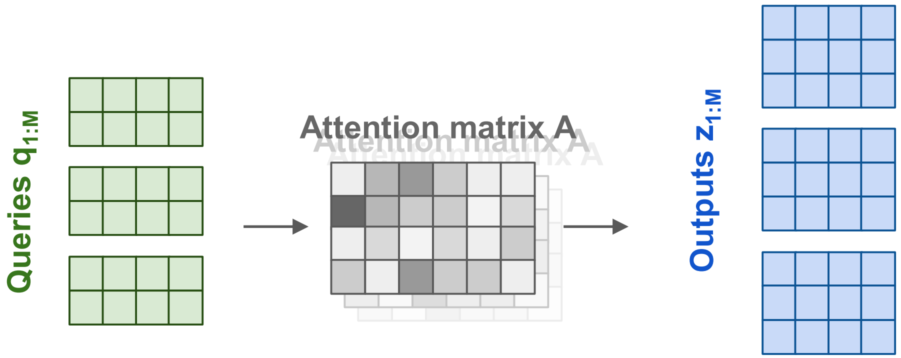
$$z_i = \begin{pmatrix} \mathrm{Attn}_1(q_i, x) \\ \mathrm{Attn}_2(q_i, x) \\ \cdots \\ \mathrm{Attn}_K(q_i, x) \end{pmatrix}$$

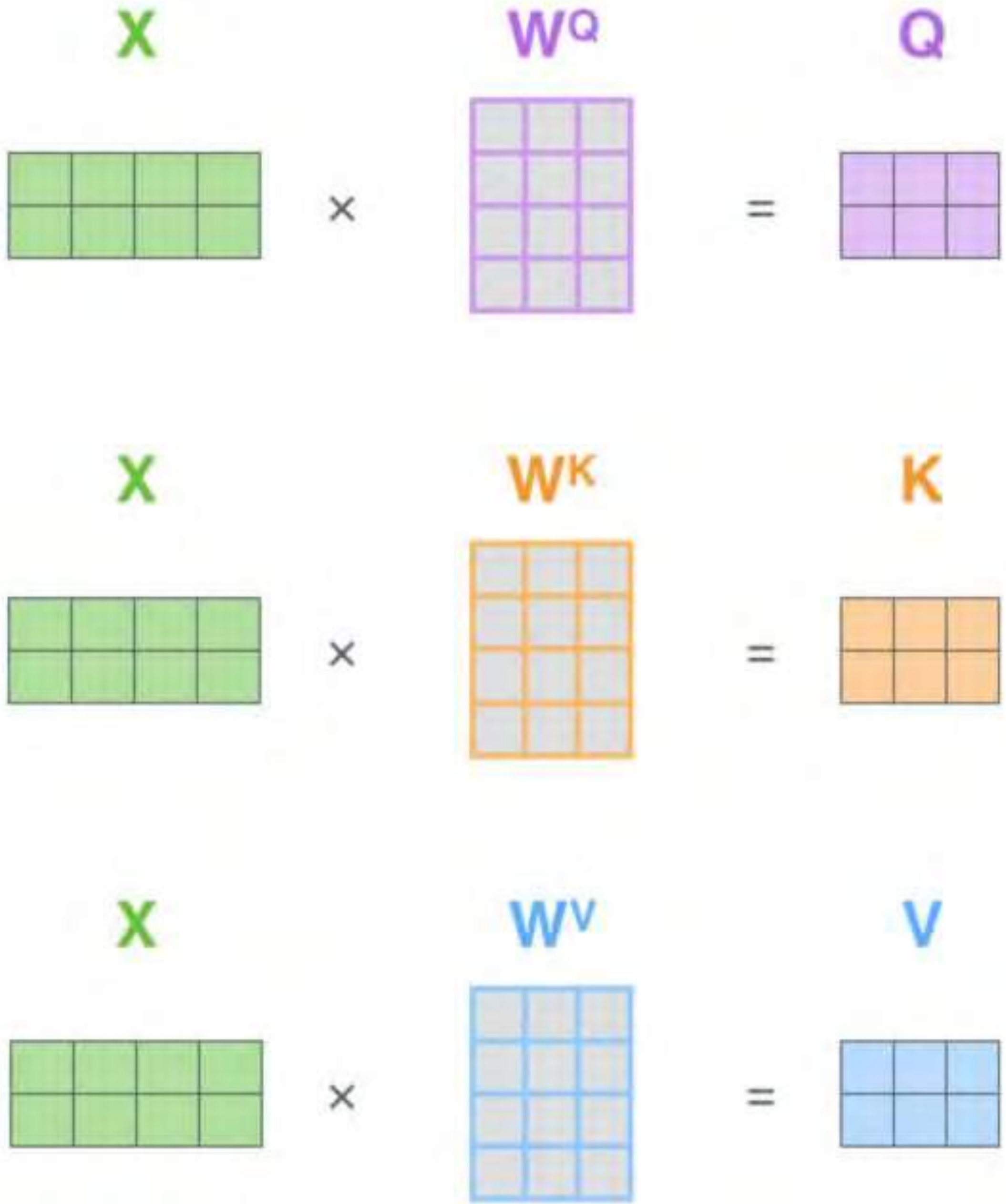


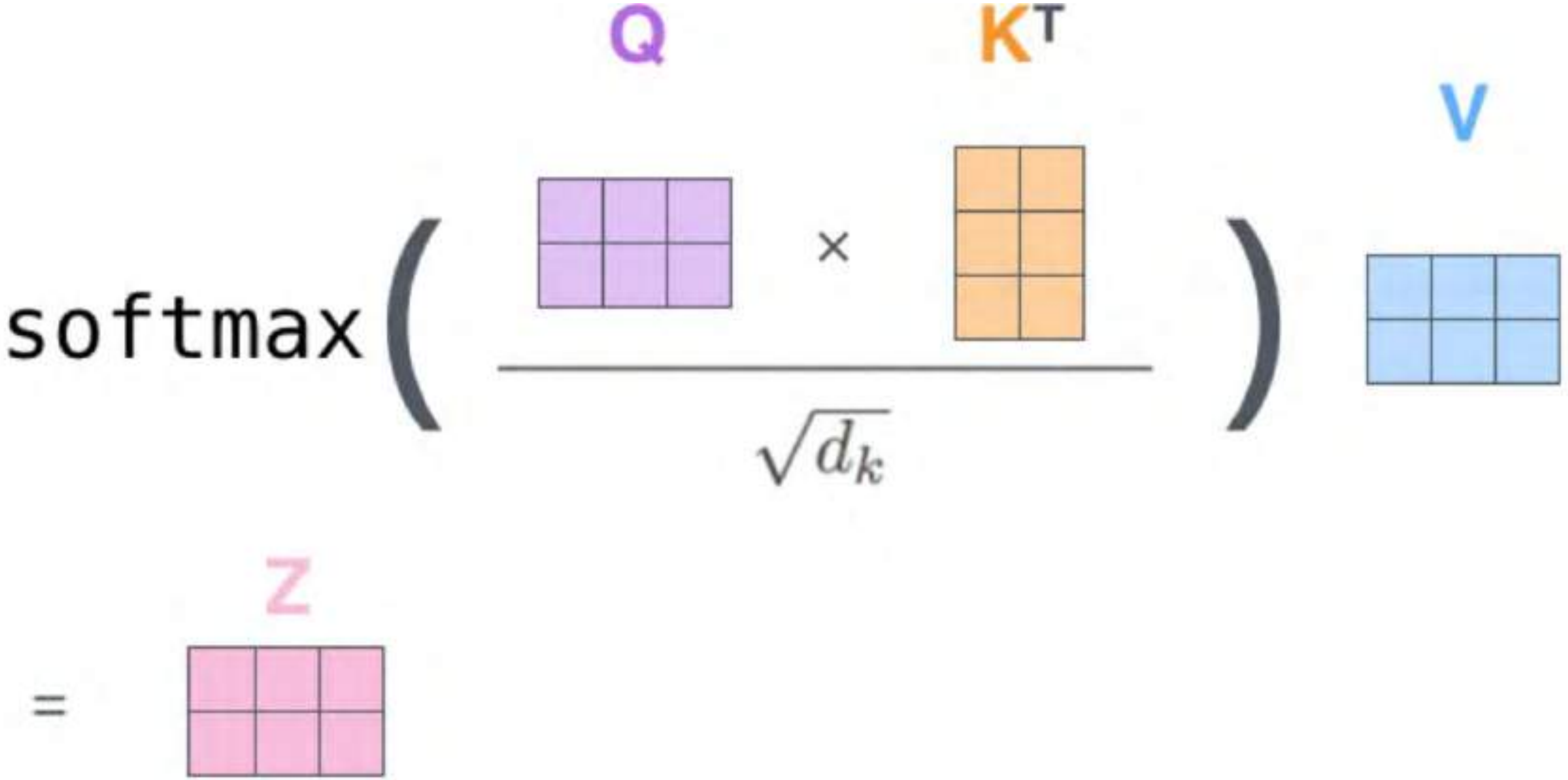
3. La presentación que resume lo anterior es:

$$z = \mathrm{softmax}(QK'/\sqrt{d_{key}})V$$

La complejidad de la ecuación es O(N²)




X
W^Q
Q
X
W^K
K
X
W^V
V



The self-attention calculation in matrix form

	"I"	"love"	"tennis"	"."
"I"	0.7	0.2	0.06	0.04
"love"	0.1	0.8	0.05	0.05
tennis"	0.05	0.1	0.75	0.1
"."	0.1	0.2	0.05	0.65

What is Self-Attention?

Self-Attention allows the model to relate words to each other.

In this simple case we consider the sequence length **seq** = 6 and $\mathbf{d}_{\mathbf{model}} = \mathbf{d}_k = 512$.

The matrices **Q**, **K** and **V** are just the input sentence.

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{softmax}\left(\frac{\underset{(6,\ 512)}{Q} \times \underset{(512,\ 6)}{K^T}}{\sqrt{512}}\right) =$$

	YOUR	CAT	IS	A	LOVELY	CAT	Σ
YOUR	0.268	0.119	0.134	0.148	0.179	0.152	1
CAT	0.124	0.278	0.201	0.128	0.154	0.115	1
IS	0.147	0.132	0.262	0.097	0.218	0.145	1
A	0.210	0.128	0.206	0.212	0.119	0.125	1
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174	1
CAT	0.195	0.114	0.203	0.103	0.157	0.229	1

(6, 6)

* a

* for simplicity I considered only one head, which makes $d_{model} = d_k$.

How to compute Self-Attention?

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.268	0.119	0.134	0.148	0.179	0.152
CAT	0.124	0.278	0.201	0.128	0.154	0.115
IS	0.147	0.132	0.262	0.097	0.218	0.145
A	0.210	0.128	0.206	0.212	0.119	0.125
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174
CAT	0.195	0.114	0.203	0.103	0.157	0.229

(6, 6)

X

V

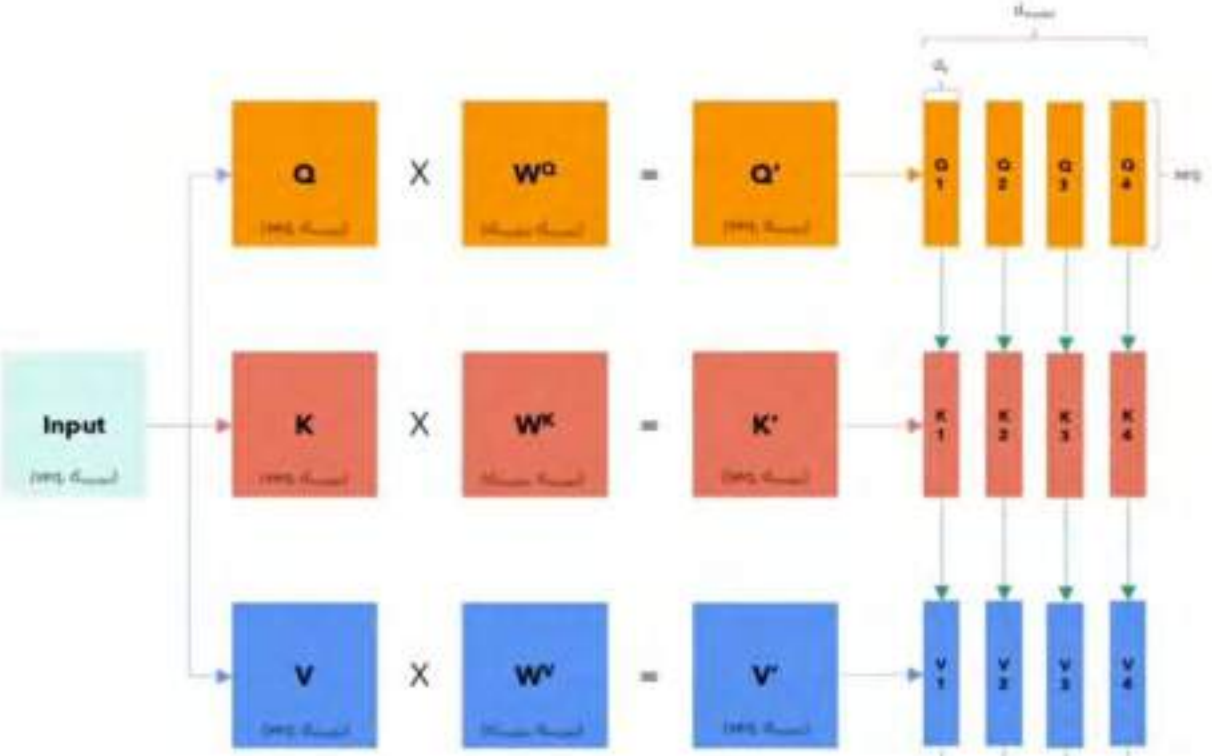
(6, 512)

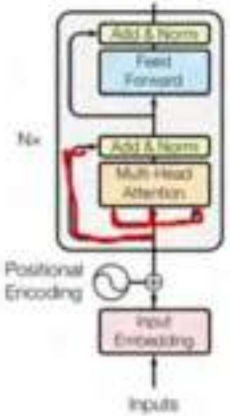
=

Attention

(6, 512)

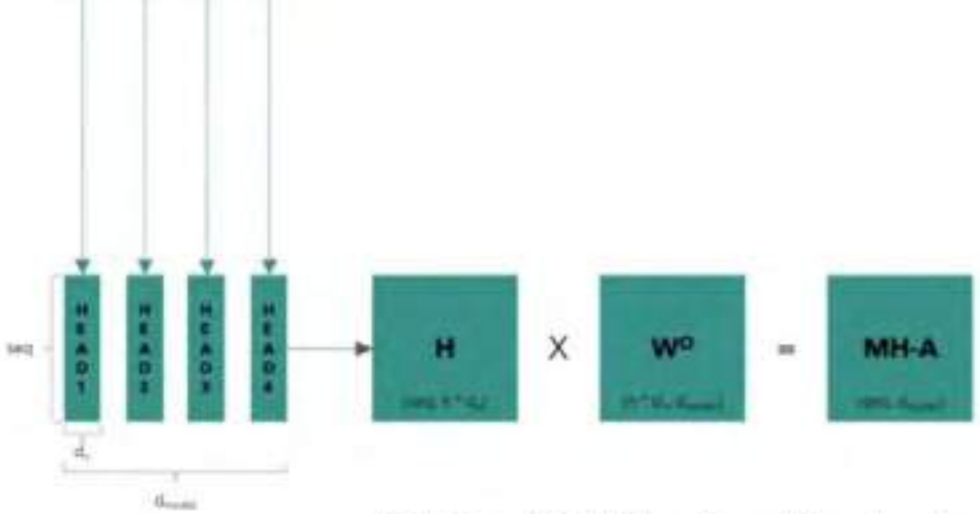
Each row in this matrix captures not only the meaning (given by the embedding) or the position in the sentence (represented by the positional encodings) but also each word's interaction with other words.





$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$



$$MultiHead(Q,K,V) = Concat(head_1 \ldots head_h)W^O$$

seq = sequence length

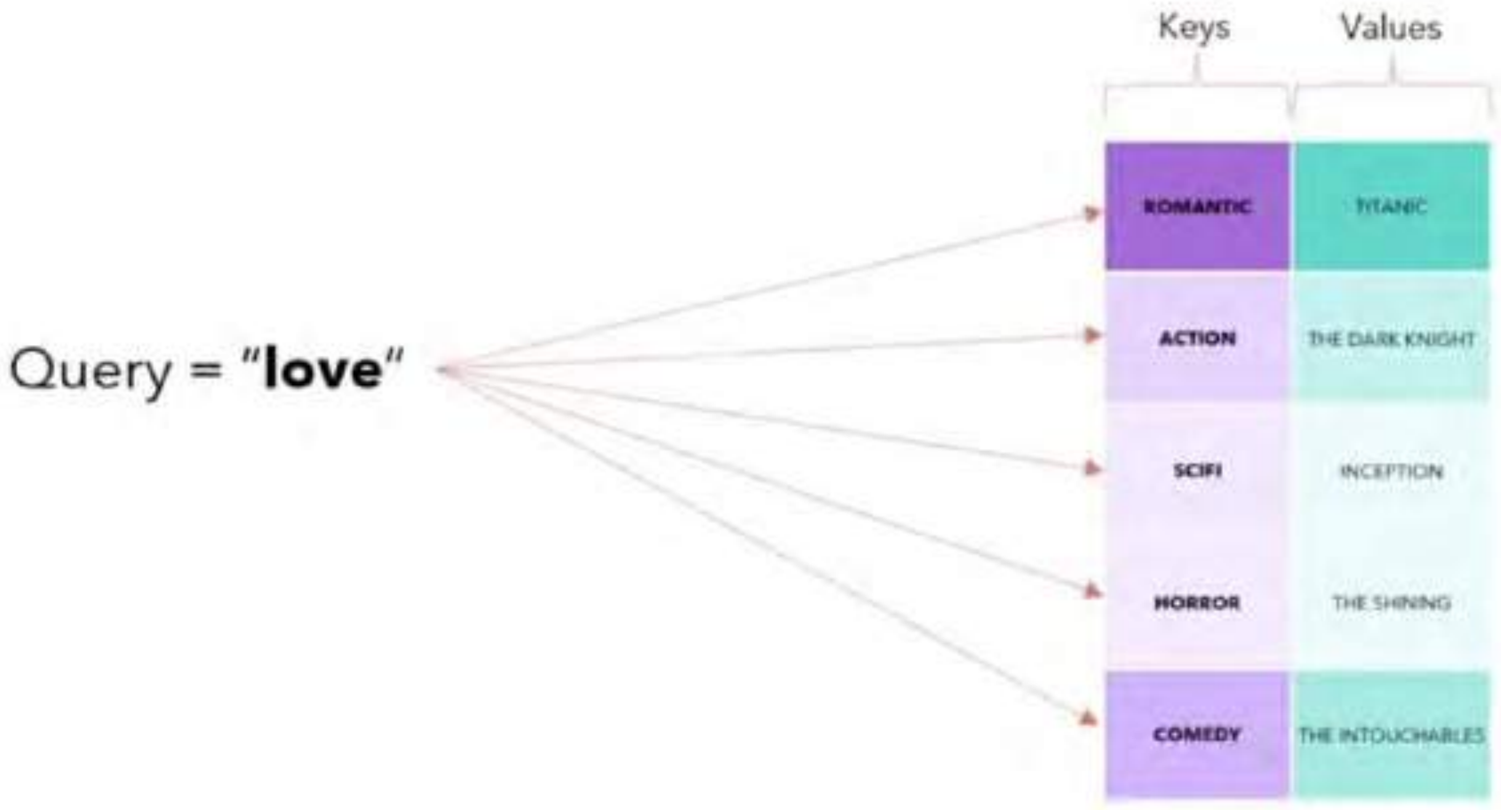
d_{model} = size of the embedding vector

h = number of heads

$d_k = d_v$ = d_{model} / h

Why query, keys and values?

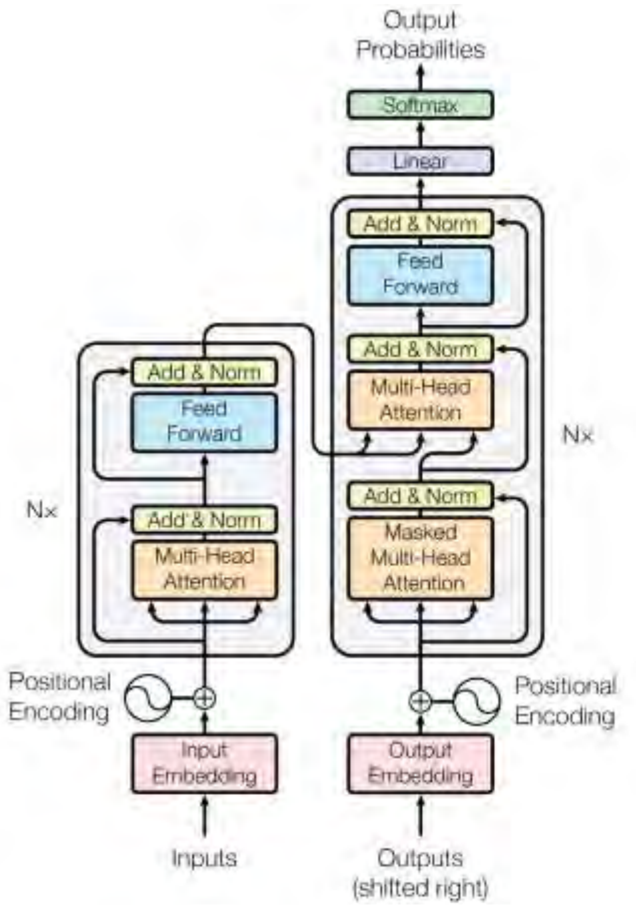
The Internet says that these terms come from the database terminology or the Python-like dictionaries.



* this could be a Python dictionary or a database table.

Attention Is All You Need - The Transformer architecture

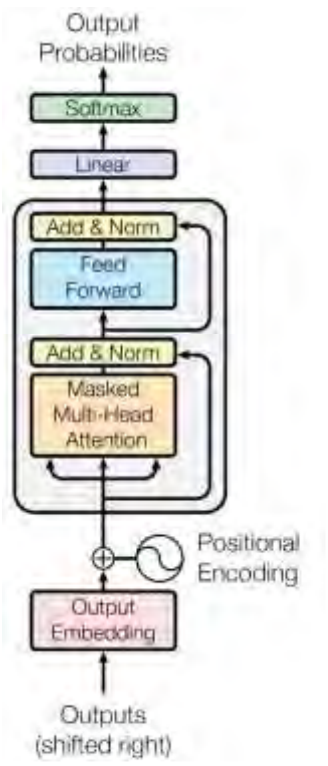
2017, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin



La primer victoria
Modelado de lenguaje / NLP

Decoder-only
GPT

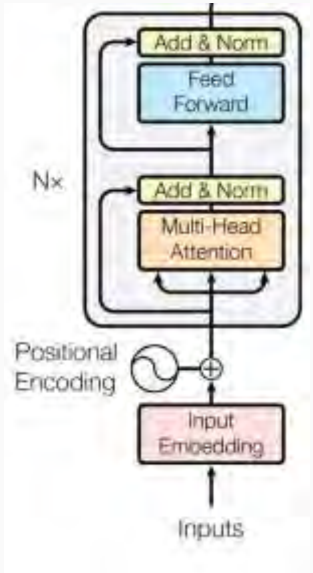
[ladra_]



[START] [El_] [perro_]

Encoder-only
BERT

[*] [*] **[ladra_]** [*] **[el_]** [*]



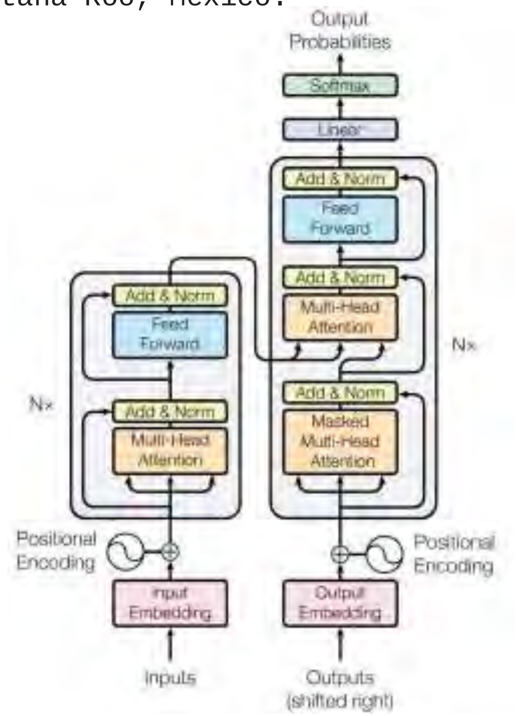
[El_] [perro_] **[MASK]** [en_] **[MASK]** [patio_]

Enc-Dec
T5

El juego se acabó.

Un grupo de personas se manifestó en la avenida …

En Quintana Roo, Mexico.



Translate EN-SPA: The game is over.

Resumir: El día de hoy en la capital …

Pregunta: ¿Donde está Cancún?

La segunda victoria:

Visión

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

2020, A Dosovitskiy, L Beyer, A Kolesnikov, D Weissenborn, X Zhai, T Unterthiner, M Dehghani, M Minderer, G Heigold, S Gelly, J Uszkoreit, N Houlsby

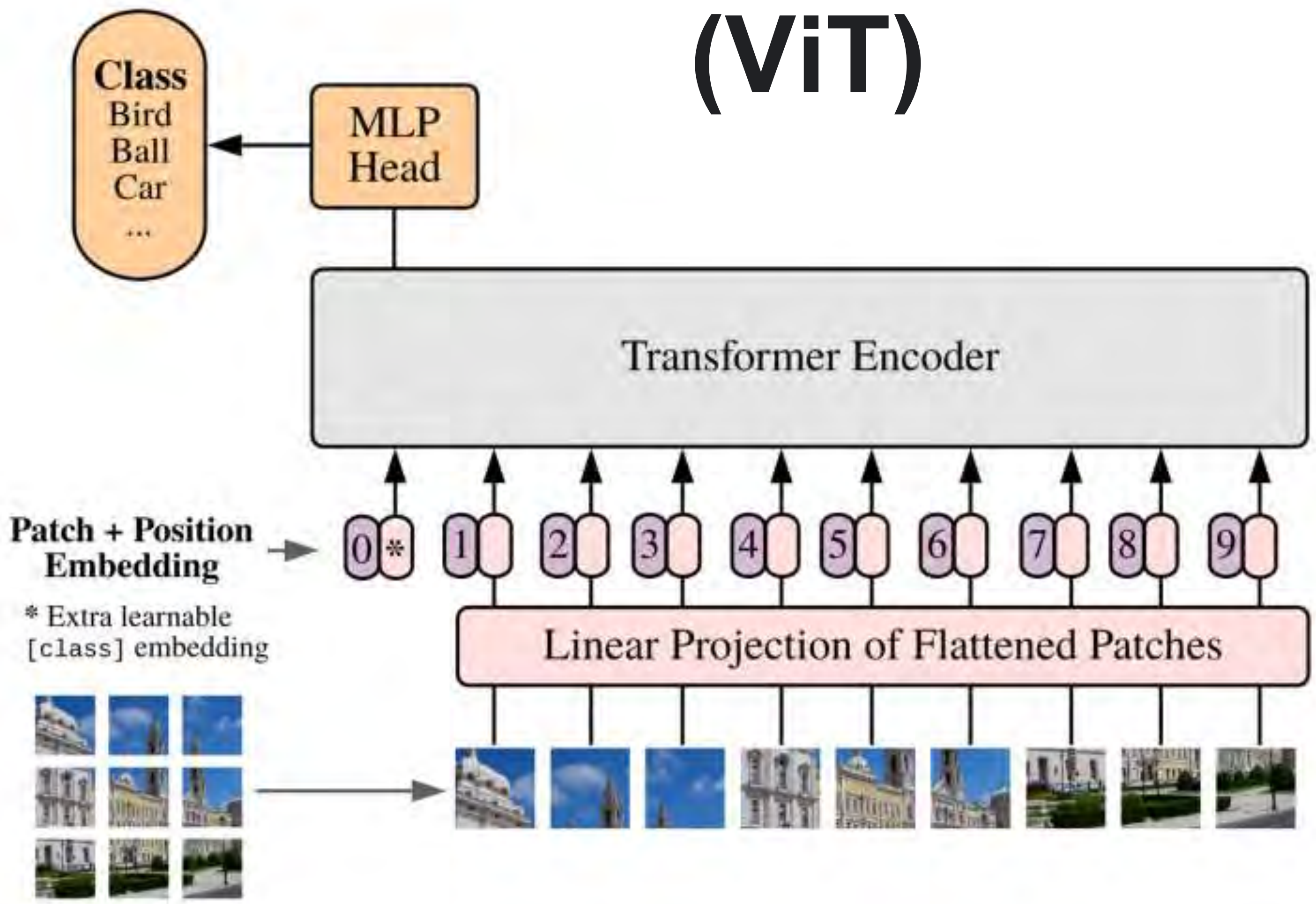
Hubo varios trabajos previos que intentaron aplicar el concepto de self-attention a nivel de pixel.

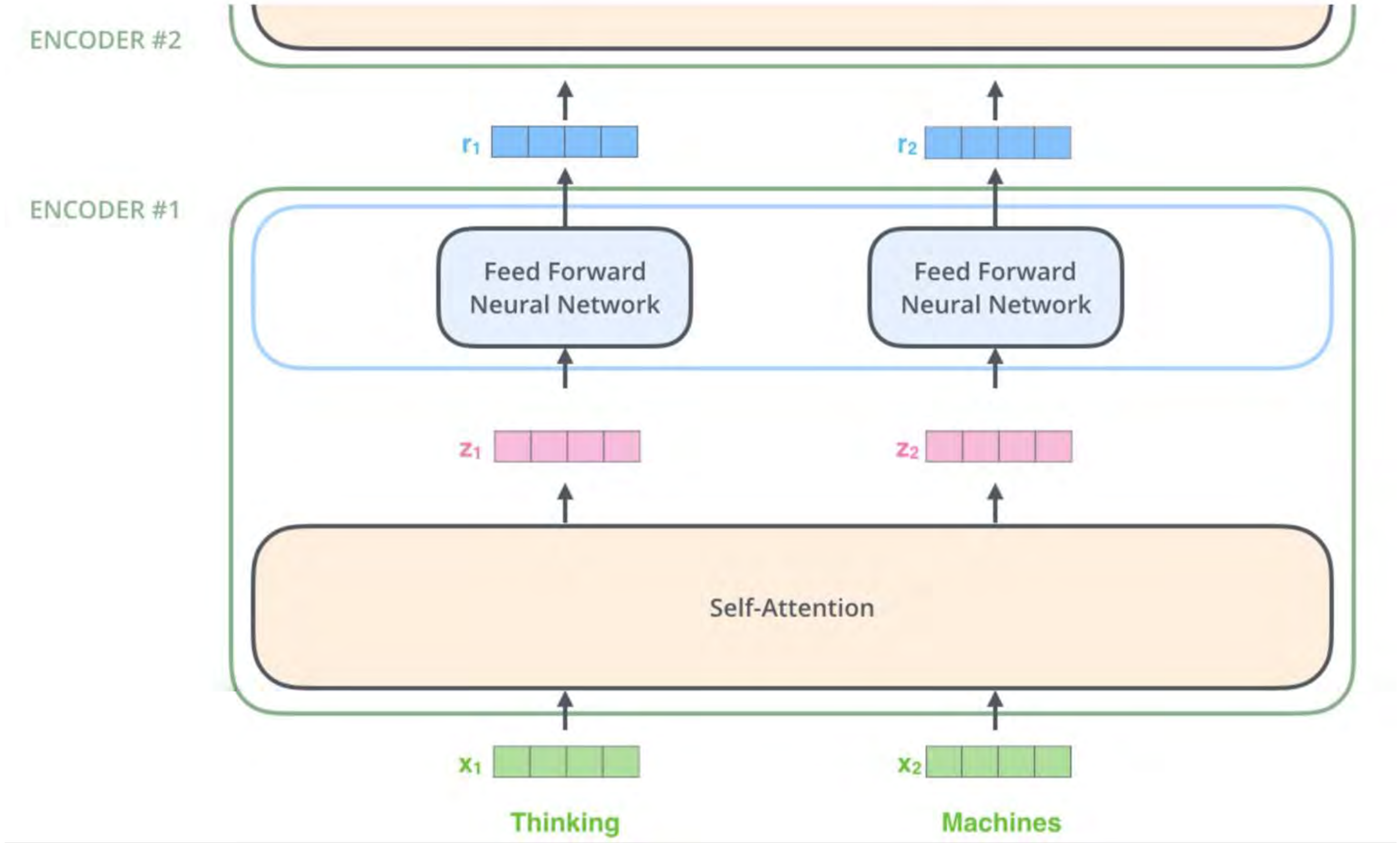
Para una imagen de 224 px², se tendría una secuencia de ~ 50 k tokens de longitud.

∴ la mayoría de trabajos restringieron aplicar bloques de atención a vecindarios locales de píxeles o bien, como un mecanismo posterior a un proceso de detección.

La **idea clave** para utilizar la arquitectura Transformer fue "tokenizar" la imagen, partiendola en **parches** de 16px² y tratar cada parche como un token. De estos tokens se obtiene su embedding y se introducen.

Vision Transformer (ViT)




ENCODER #2
r_1
r_2
ENCODER #1
Feed Forward Neural Network
Feed Forward Neural Network
z_1
z_2
Self-Attention
x_1
x_2
Thinking
Machines

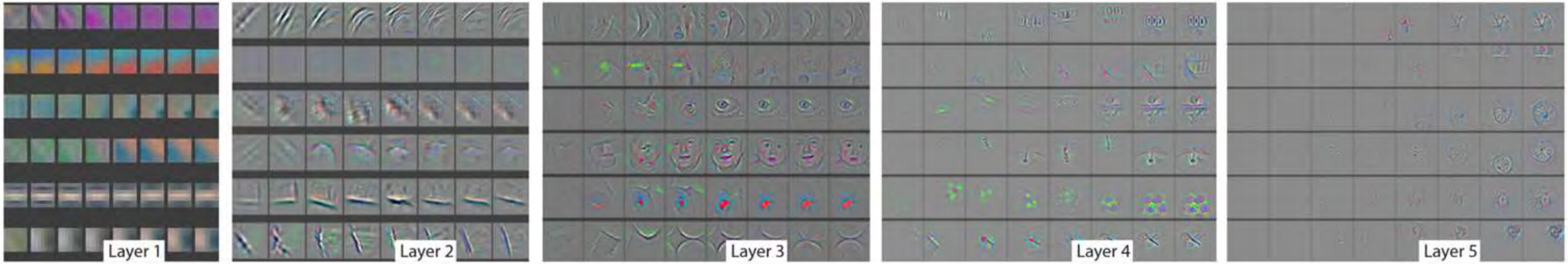
❷ **Do MSAs act like Convs?** We show that MSAs and Convs exhibit opposite behaviors. MSAs aggregate feature maps, but Convs diversify them. Moreover, as shown in Fig. 2a, the Fourier analysis of feature maps shows that MSAs reduce high-frequency signals, while Convs, conversely, amplifies high-frequency components. In other words, *MSAs are low-pass filters, but Convs are high-pass filters*. In addition, Fig. 2b indicates that Convs are vulnerable to high-frequency noise but that MSAs are not. Therefore, MSAs and Convs are complementary.

As expected, the result in Fig. 2b reveals that ViT and ResNet are vulnerable to low-frequency noise and high-frequency noise, respectively. Low-frequency signals and the high-frequency signals each correspond to the shape and the texture of images. The results thus suggests that MSAs are shape-biased (Naseer et al., 2021), whereas Convs are texture-biased (Geirhos et al., 2019).

Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler ZEILER@CS.NYU.EDU
Dept. of Computer Science, Courant Institute, New York University

Rob Fergus FERGUS@CS.NYU.EDU
Dept. of Computer Science, Courant Institute, New York University



2013

What do Vision Transformers Learn? A Visual Exploration

2022

Amin Ghiasi[*1] **Hamid Kazemi**[*1] **Eitan Borgnia**[1] **Steven Reich**[1] **Manli Shu**[1]

Micah Goldblum[2] **Andrew Gordon Wilson**[2] **Tom Goldstein**[1]
[1] University of Maryland - College Park [2] New York University [*] Equal contribution

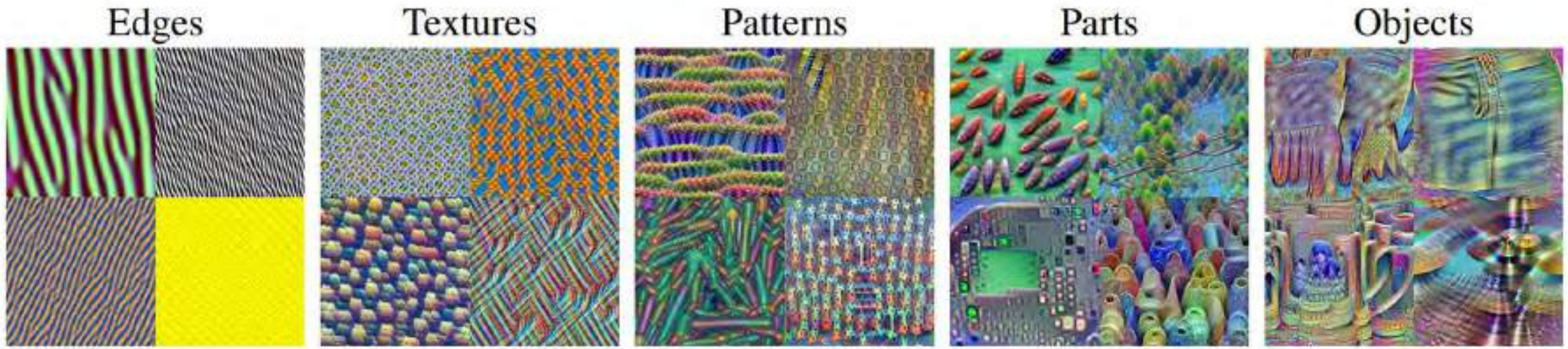


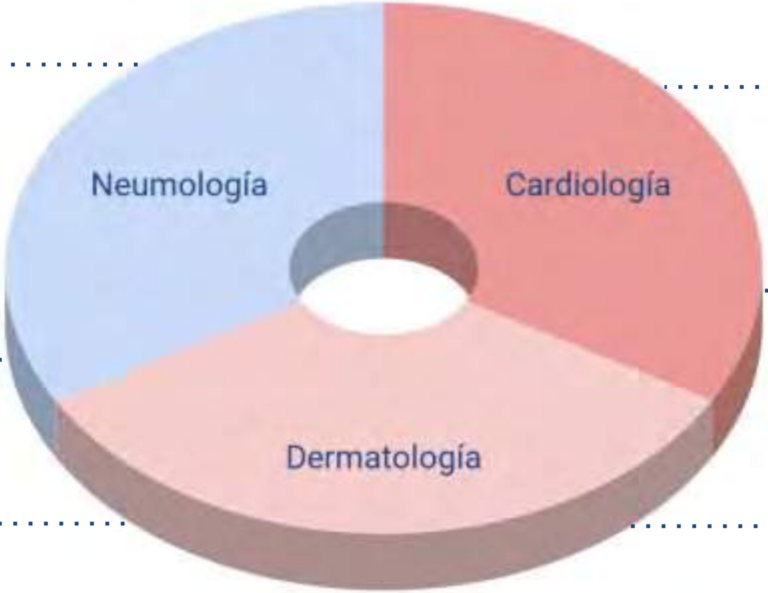
Figure 1: **The progression for visualized features of ViT B-32.** Features from early layers capture general edges and textures. Moving into deeper layers, features evolve to capture more specialized image components and finally concrete objects.

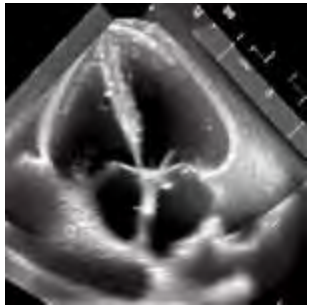
Stable diffusion XL

Prompt: **Emma Watson as a powerful mysterious sorceress, casting lightning magic, detailed clothing**

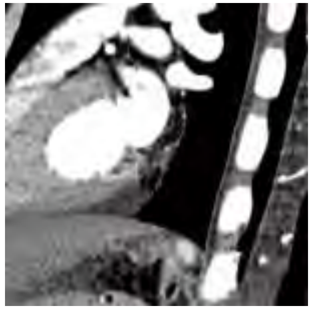
Generación de imágenes médicas artificiales con modelos de difusión

Imágenes artificiales de neumonía por COVID-19





Imágenes artificiales de tomografía y ecocardiografía de corazón

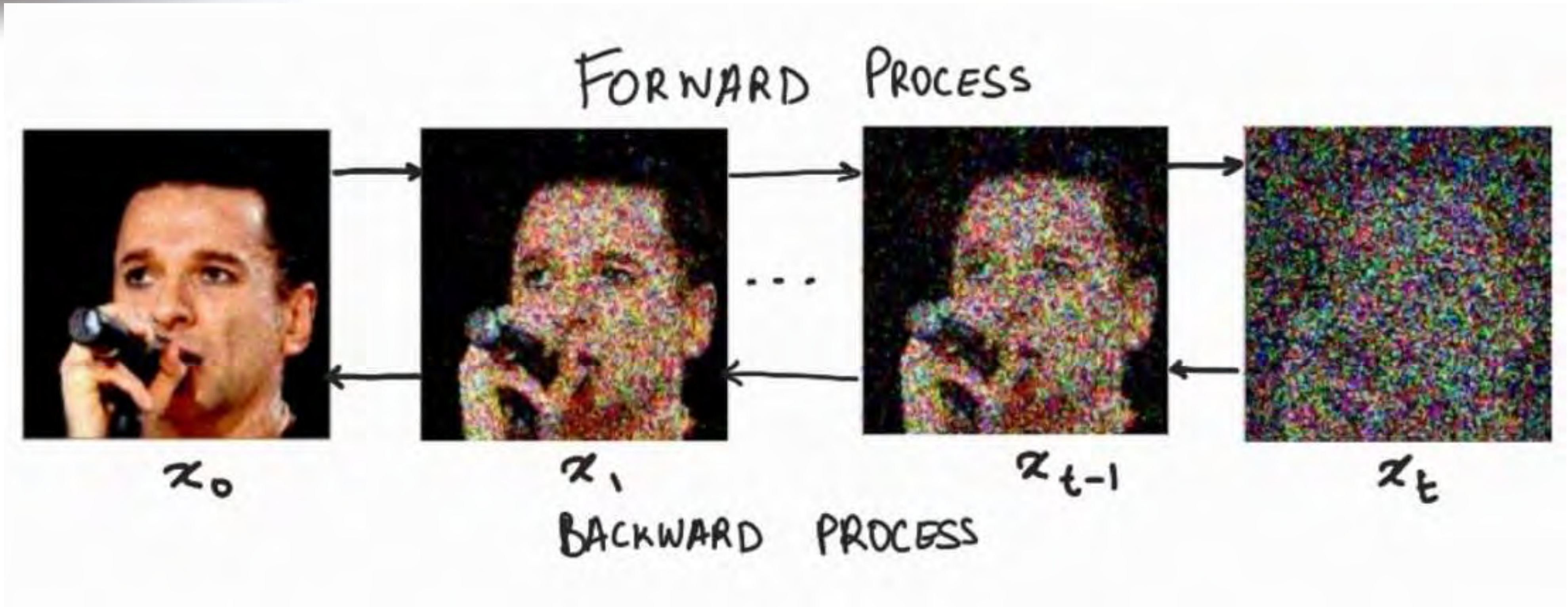


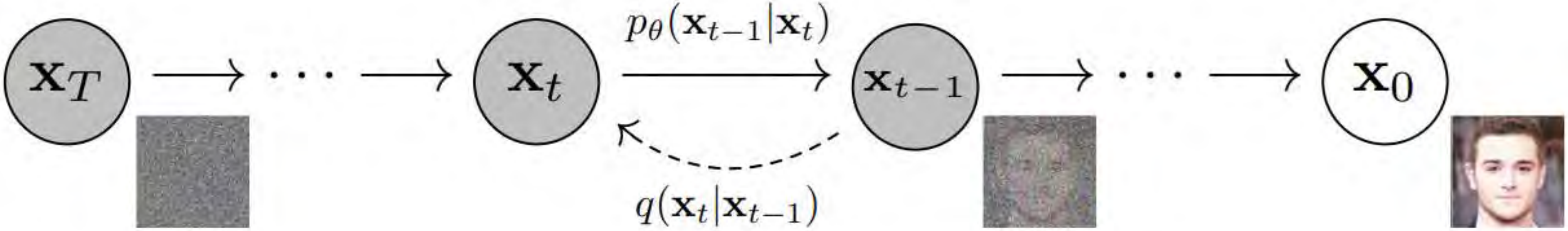
Buscamos compensar la falta de datos disponibles en el área médica 💊 generando datos artificiales.

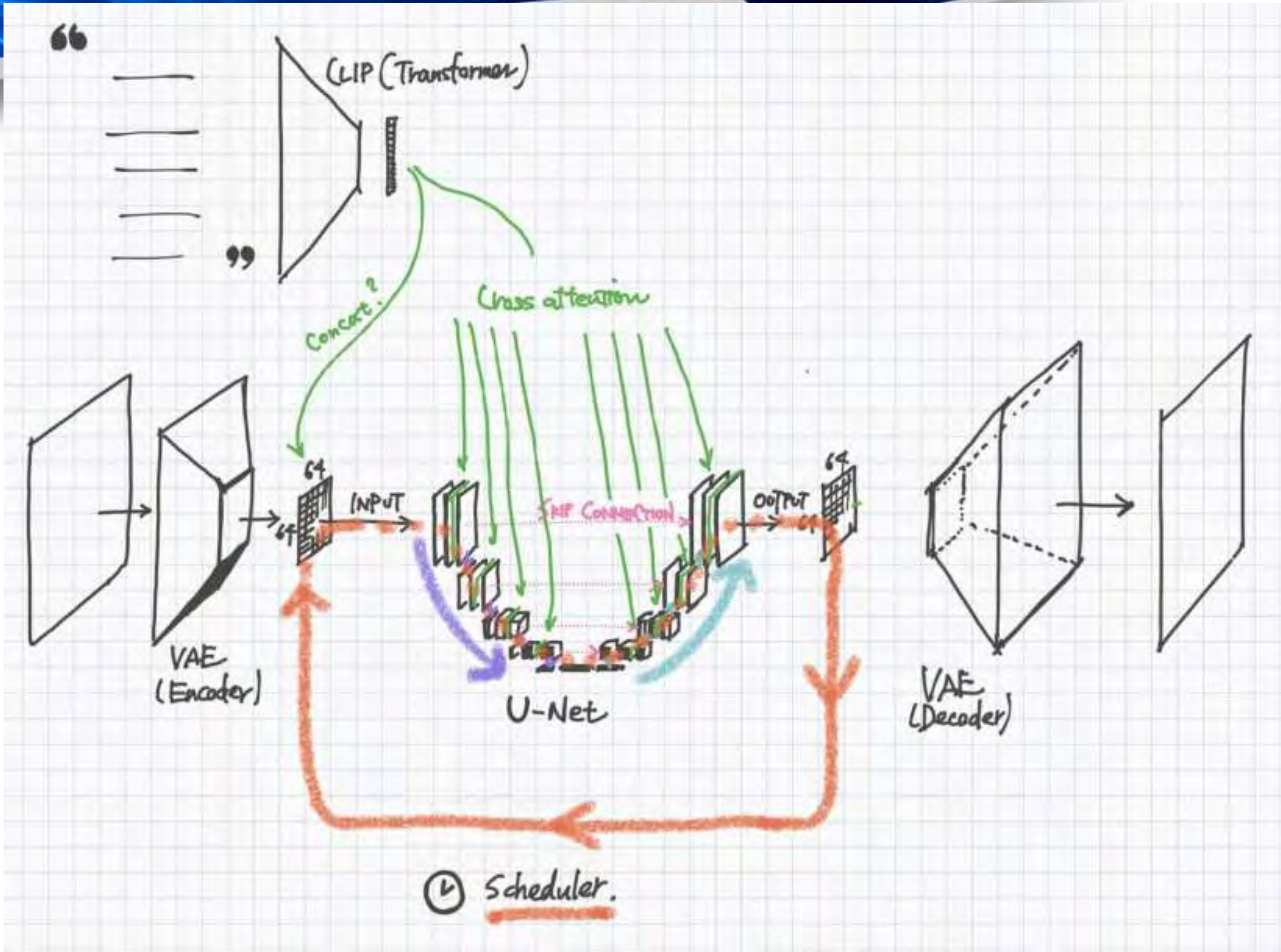
Imágenes artificiales de melanoma (cáncer de piel)

Queremos entrenar redes neuronales capaces de identificar diferentes afecciones 🔬 .

Denoising Diffusion Probabilistic Models (DDPM)





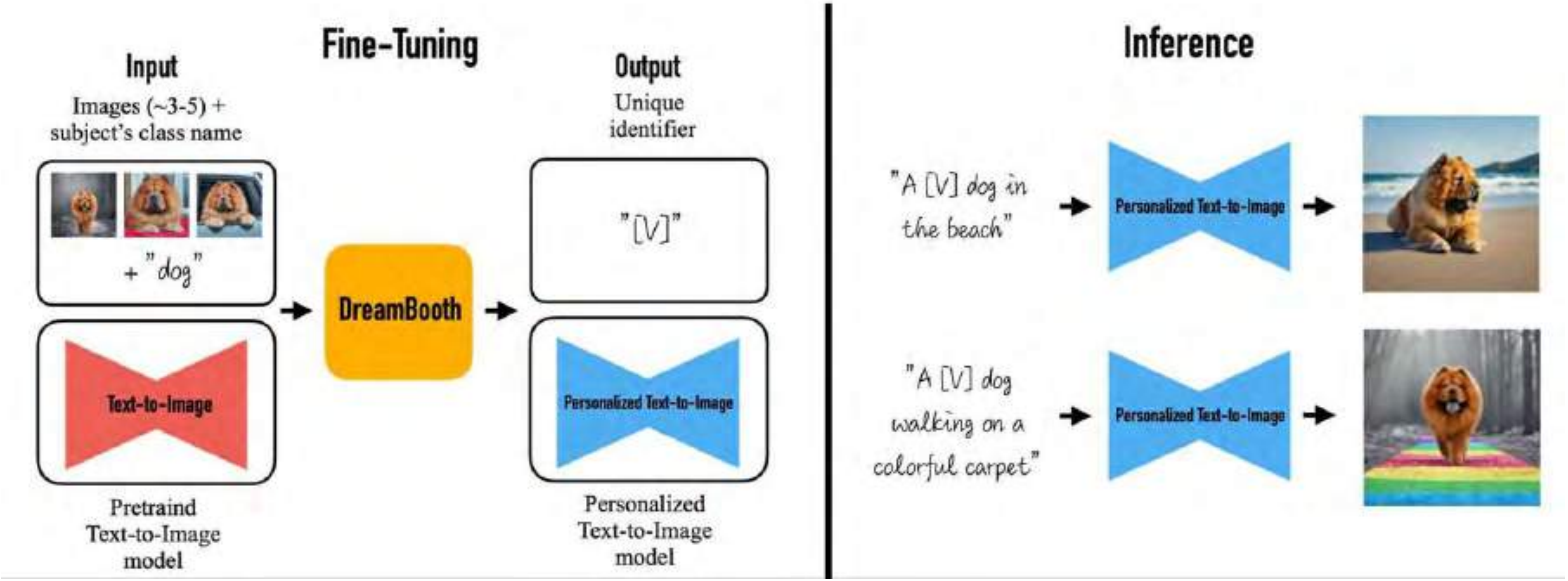


CLIP (Transformer)
Concat?
Cross attention
64
64
INPUT
SKIP CONNECTION
OUTPUT
64
VAE
(Encoder)
U-Net
VAE
(Decoder)
Scheduler.

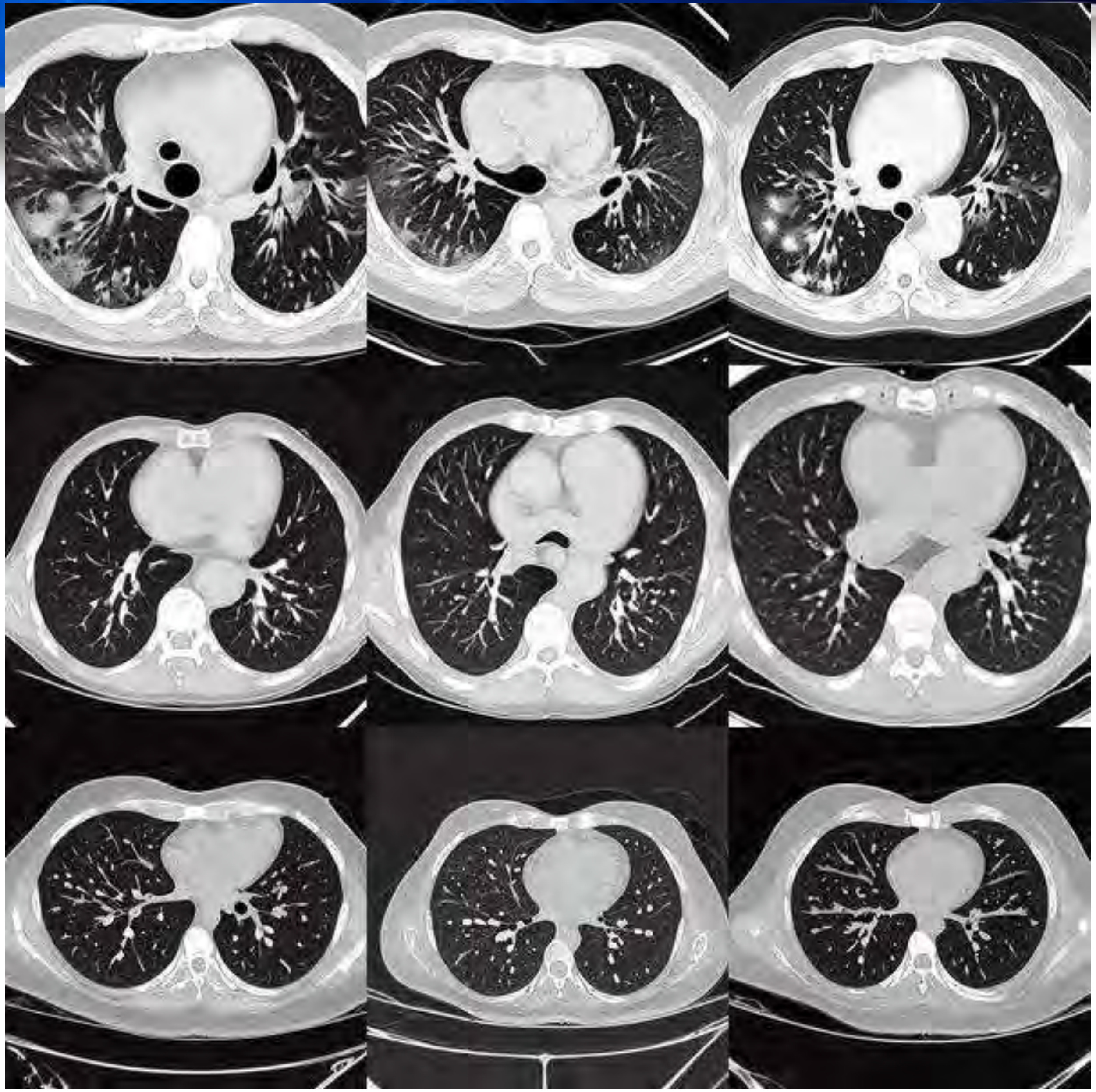
Dreambooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

"Personalización" de un modelo de difusión

Dreambooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation



Funcionamiento de Dreambooth

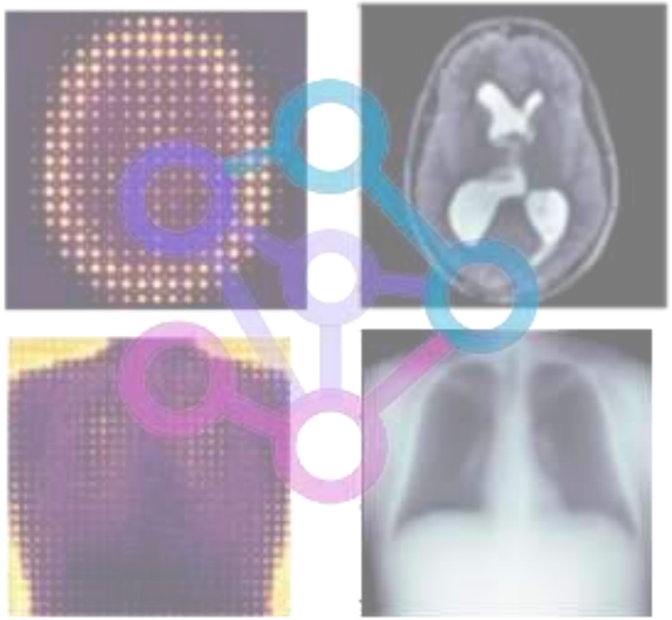


Neumología

El primer renglón son imágenes de neumonía por COVID-19, el segundo renglón son imágenes de neumonía por causas diferentes a COVID-19 y el tercer renglón son imágenes de pacientes normales

Se generaron 1,000 imágenes sintéticas por clase.

PumaMedNet
Plataforma de IA generativa para el análisis explicable de diagnósticos con imágenes médicas



Carlos Minutti-Martinez[1], Boris Escalante-Ramirez[2], Jimena Olveres[2]

[1]Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación (INFOTEC)

[2]Universidad Nacional Autónoma de México (UNAM)

2026

Diagrama esquemático

Arquitectura base

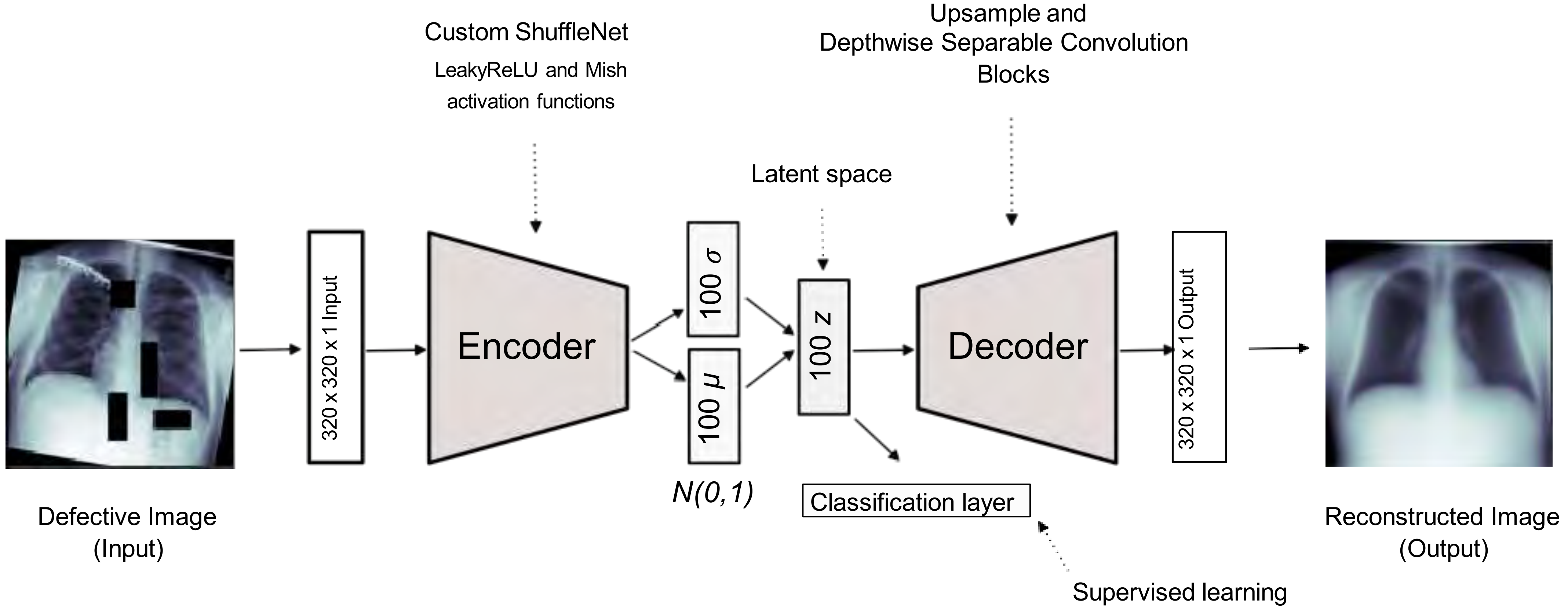


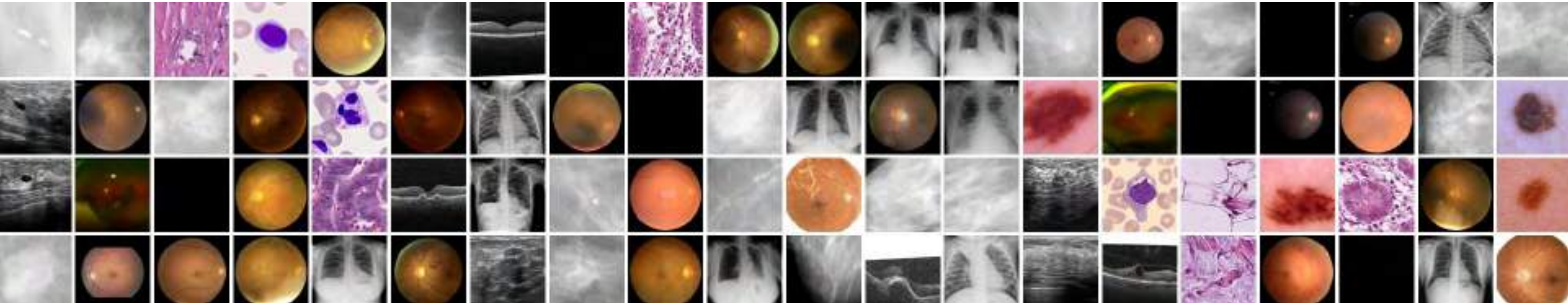
Figura: Diagrama esquemático de la arquitectura del modelo propuesto.

Pre-entrenamiento
Dominio de imágenes médicas

Pre-entrenamiento **supervisado** y **no supervisado**

> 17 conjuntos de datos públicos
(microscopía de células sanguíneas, radiografías de tórax, ultrasonido, mamografías, histopatología, H&E colorrectal, dermatoscópicas, fotografías de fondo de ojo, OCT)
> 28 tareas de clasificación
> 372.895 imágenes

Permite capturar características y patrones visuales propios del dominio médico.



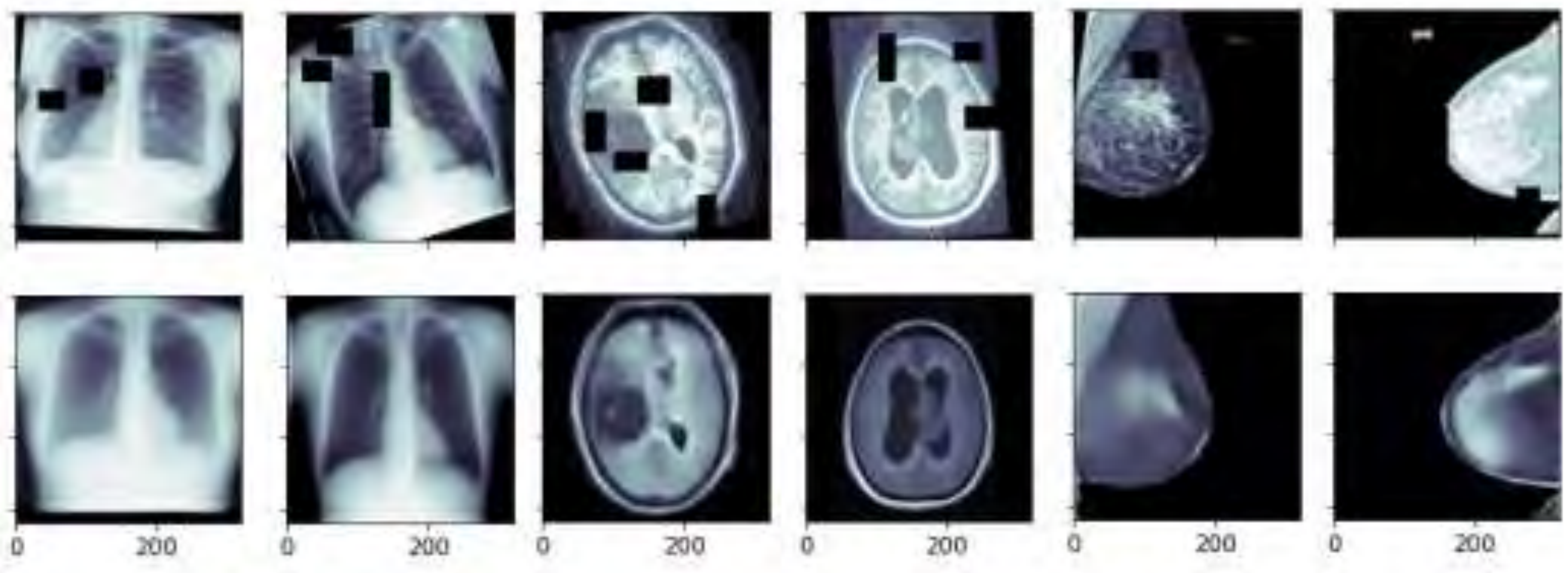
Mecanismos de contexto

Interpretabilidad con IA generativa

Diseño inspirado en arquitecturas tipo *Transformer*, incorporando mecanismos de **contexto** y **atención**.

Contexto: Aumentado extensivo de datos.

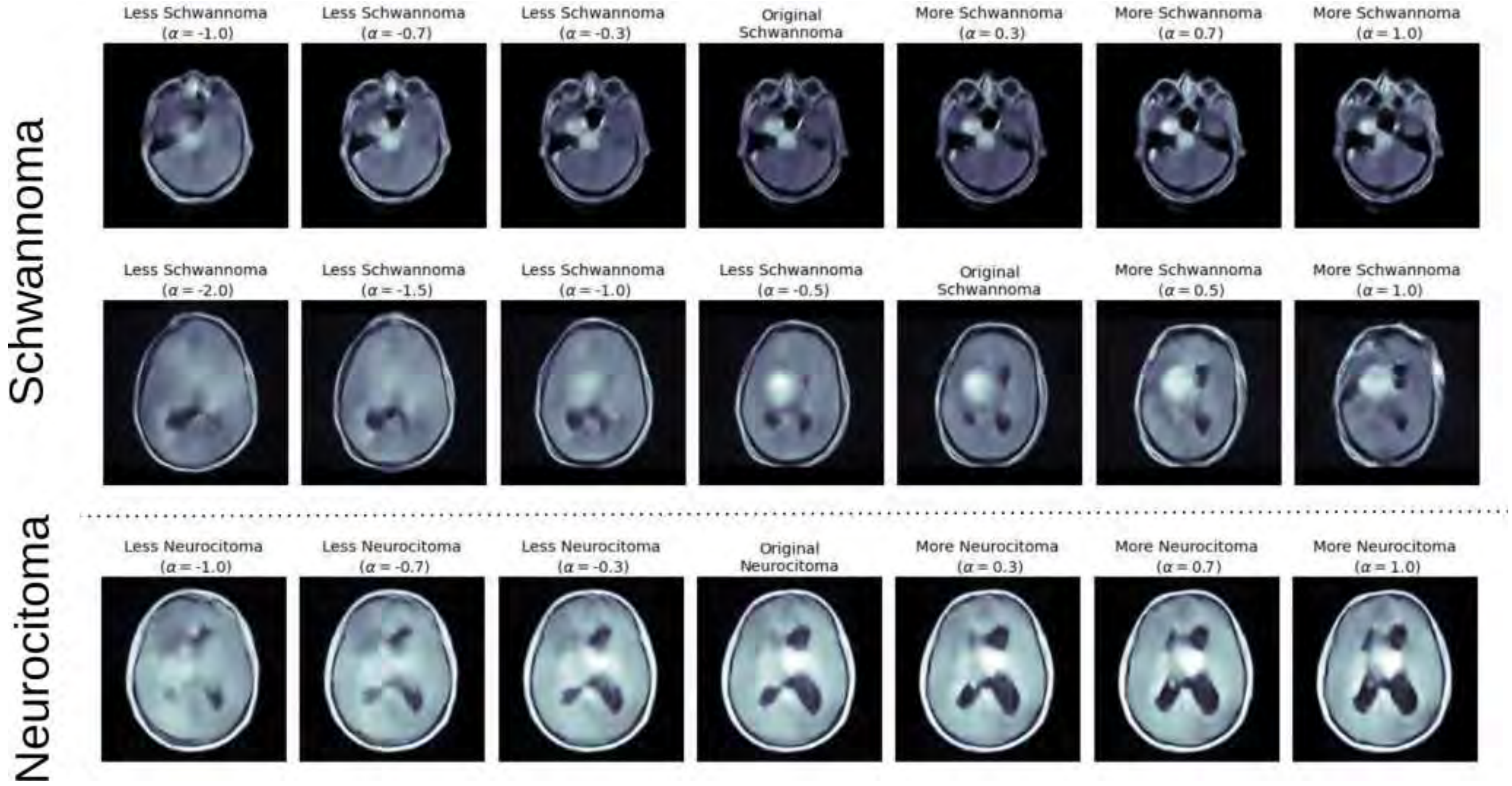
Rotaciones aleatorias, volteos, desenfoque, transformaciones de perspectiva y **borrado aleatorio** (análogo al enmascaramiento de tokens en modelos de lenguaje).



Contexto mediante la **predicción de parches faltantes** o defectuosos.

Explicabilidad

Interpretabilidad con IA generativa



De manera similar, en la presente imagen se marca toda un área contrastada correspondiente a la Efusión.

Radiografía de Tórax

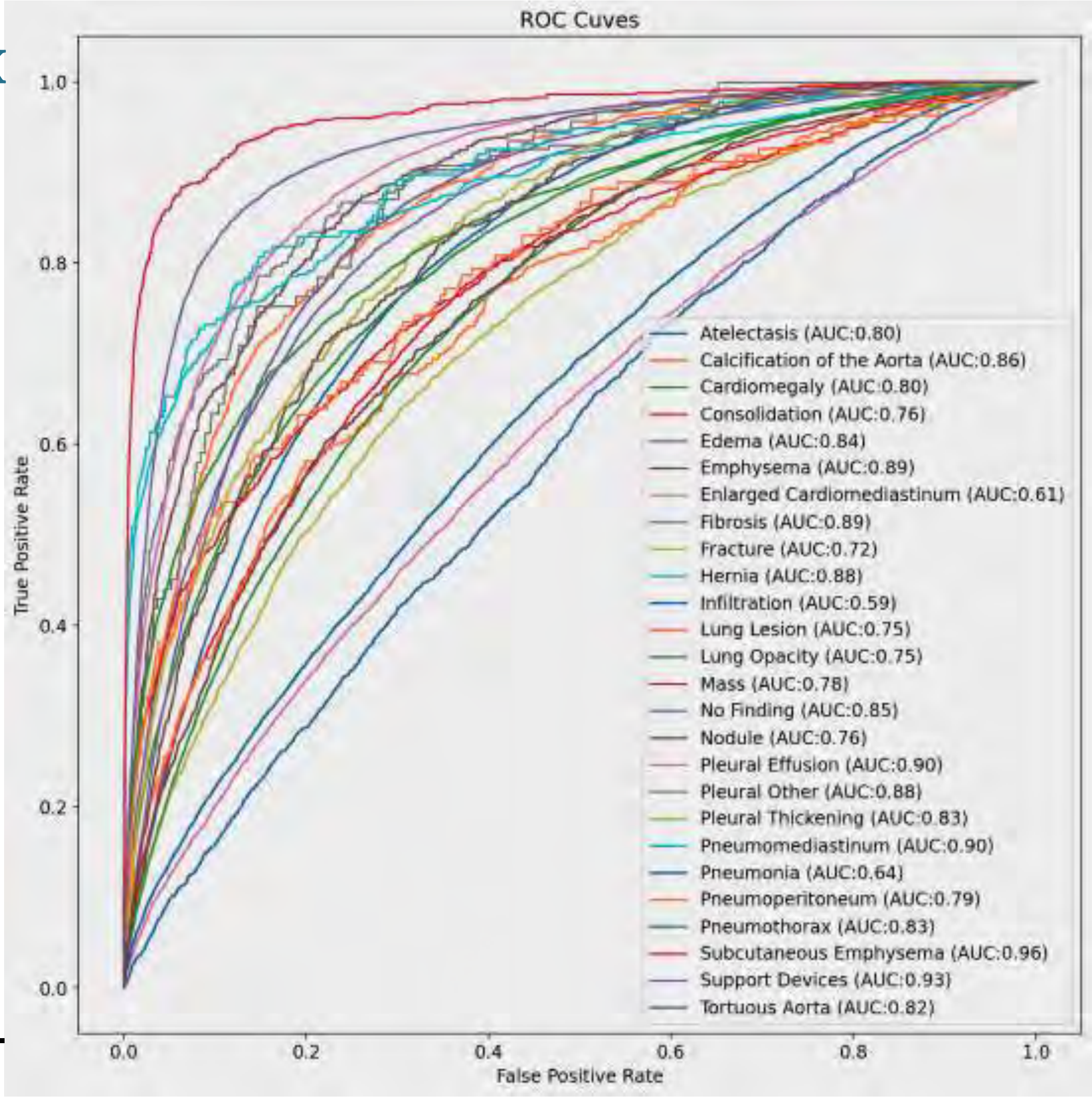
MIMIC-CXR-LT es una base de datos pública y ampliamente utilizada que contiene radiografías de tórax, junto con informes de radiología en texto libre.

+ 243 mil imágenes

+ 65 mil pacientes

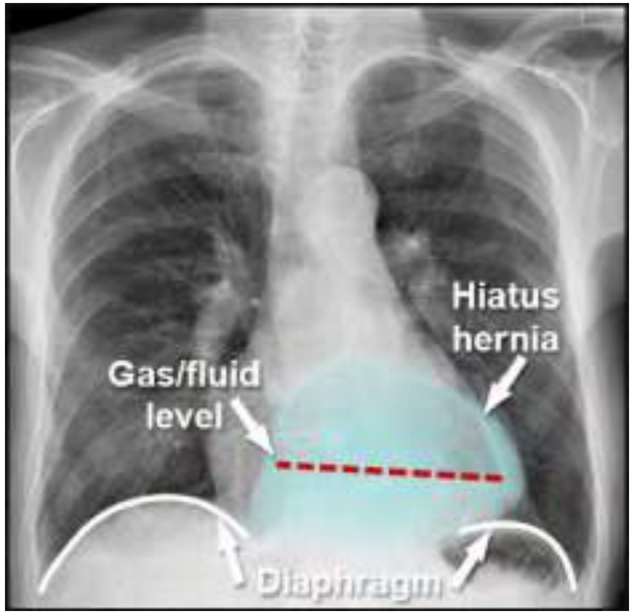
+ 26 categorías

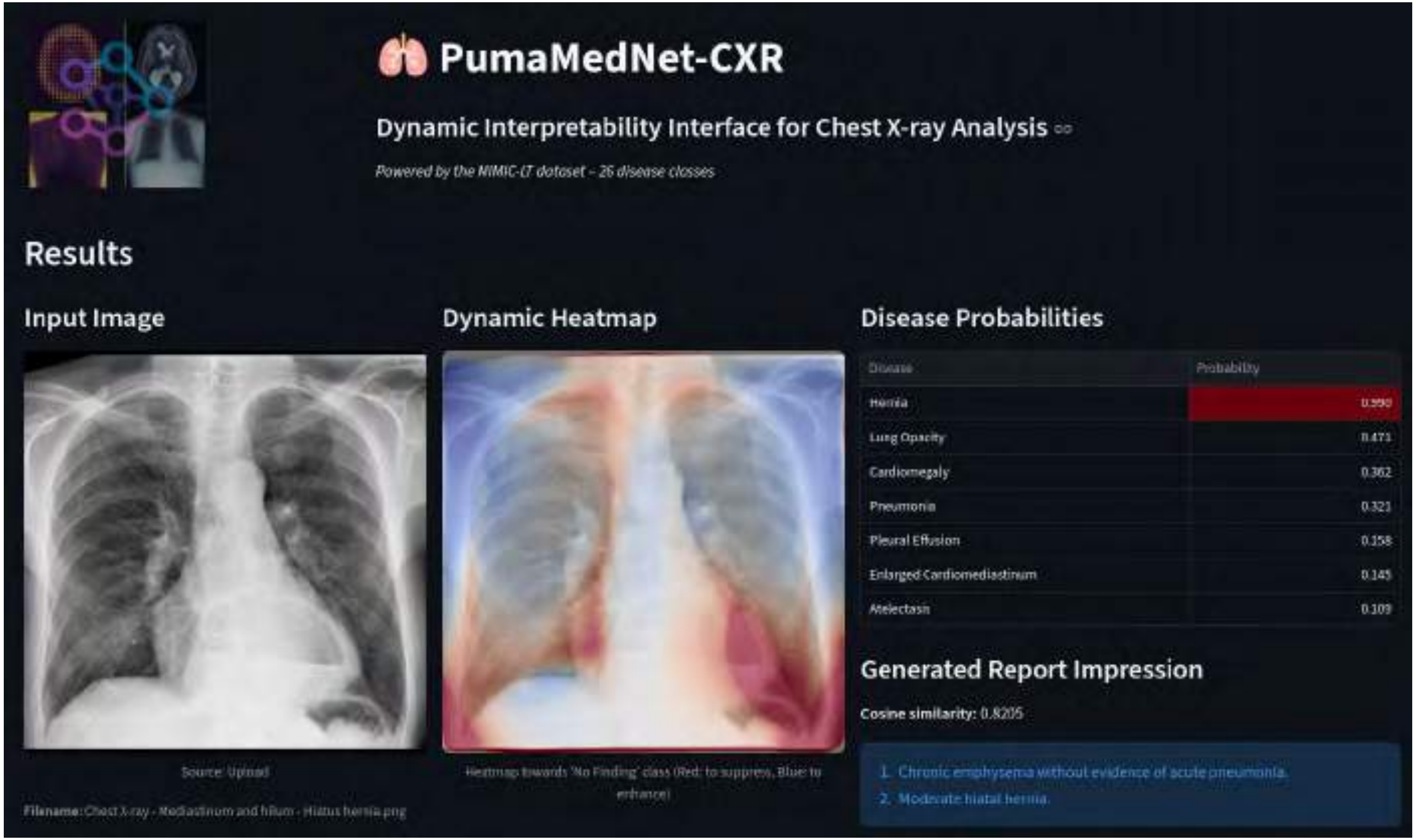
AUC promedio: **0.80**
Parámetros: **1.89M** vs **11.18M** (ResNet-18)
5.9× menor complejidad con desempeño equivalente.



https://cminuttim.github.io/PumaMedNet-CXR/

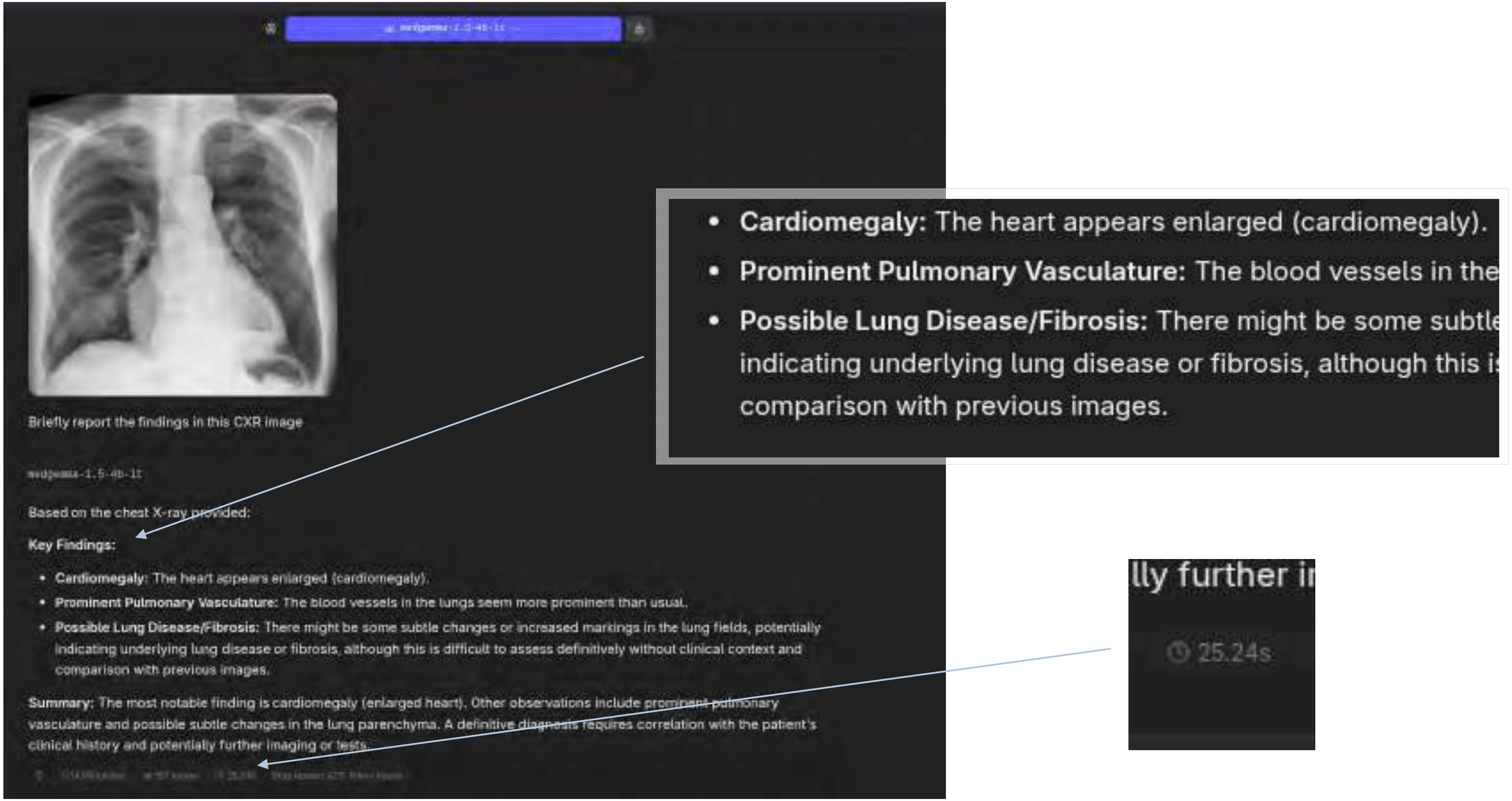
PumaMedNet → Hernia





https://www.radiologymasterclass.co.uk/gallery/chest/mediastinum_hilum/hiatus_hernia

MedGemma → Cardiomegalia



Reconocimientos

2023

Best Paper Award MICAI

Primer Lugar

2024

Premio AFIRME-FUNAM

4ª Edición

Tercer Lugar

2025

Top 10 en el programa **"Acelerando México con Inteligencia Artificial"** de **Intel**

TLR 4-5

Conclusiones

GRACIAS!

http://lapi.unam.mx/
boris@unam.mx

LaPI
LABORATORIO AVANZADO DE PROCESAMIENTO DE IMÁGENES