

Universidad de Granada

Segmentación de cáncer de próstata en imágenes IRM potenciadas en T2 usando modelos de deep learning

Trabajo Fin de Grado

Pilar Navarro Ramírez

Tutores: Francisco Herrera Triguero
Julián Luengo Martín

7 de septiembre de 2022

Descripción del problema

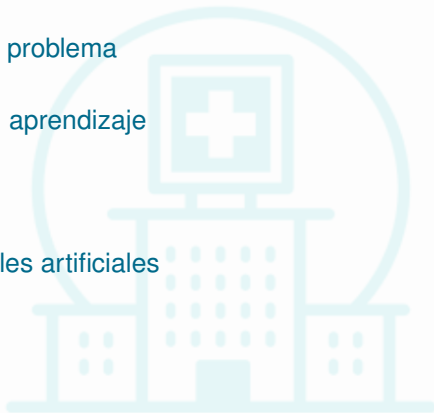
El problema del aprendizaje

Optimización

Redes neuronales artificiales

Aplicación

Trabajo Futuro



Descripción del problema

El problema del aprendizaje

Optimización

Redes neuronales artificiales

Aplicación

Trabajo Futuro



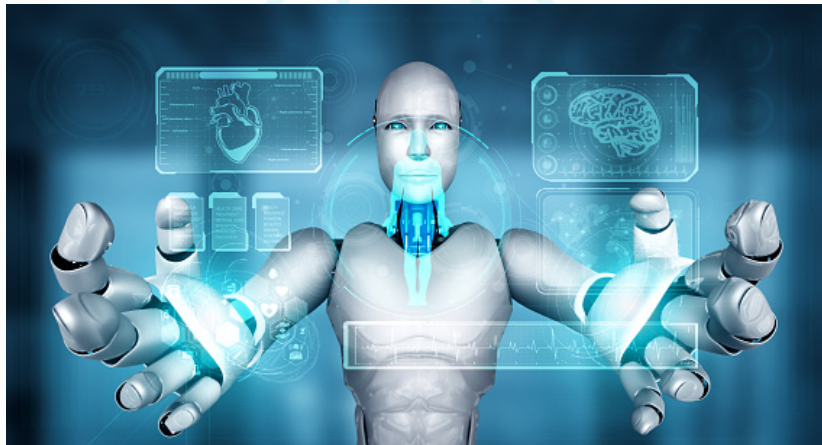
- ▶ El cáncer de próstata (CaP) es el tipo de cáncer diagnosticado con más frecuencia entre los hombres alrededor del mundo y el quinto en número de muertes.
- ▶ Una detección temprana puede hacer que la enfermedad se pueda tratar y disminuir la tasa de mortalidad.
- ▶ El uso de **imagen por resonancia magnética** (IRM) ha demostrado ser muy eficaz para diagnosticar el cáncer en una etapa temprana.

Sin embargo la interpretación de IRM ...

- ▶ Requiere de un radiólogo experto
- ▶ Es una tarea compleja en la que es necesario invertir mucho tiempo y esfuerzo
- ▶ Está sujeta a errores debidos a las limitaciones humanas

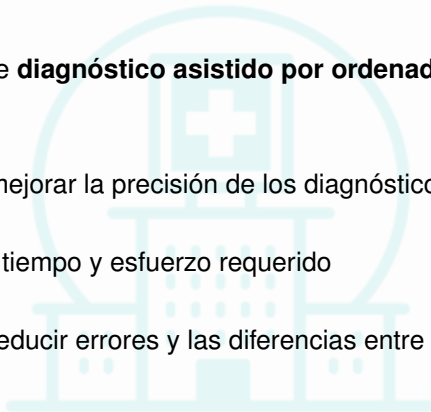


¡Inteligencia artificial al rescate!



Los sistemas de **diagnóstico asistido por ordenador** (DAO)

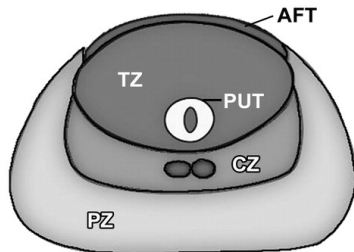
- ▶ ayudan a mejorar la precisión de los diagnósticos
- ▶ reducen el tiempo y esfuerzo requerido
- ▶ ayudan a reducir errores y las diferencias entre interpretaciones.



La próstata se divide en varias zonas bien diferenciadas:

- ▶ El estroma fibromuscular anterior (AFT)
- ▶ La zona de transición (TZ)
- ▶ La zona central (CZ)
- ▶ La zona periférica (PZ)

$PZ + TZ = CG$ (glándula central)



La **segmentación semántica** es un problema de visión por computador que consiste en etiquetar cada píxel de una imagen con una cierta clase, según lo que muestre.



Figura: Imagen original

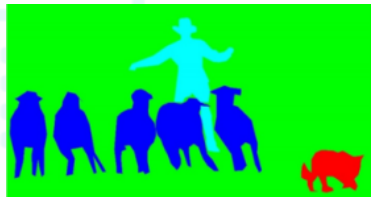


Figura: Imagen segmentada

Descripción del problema

El problema del aprendizaje

Optimización

Redes neuronales artificiales

Aplicación

Trabajo Futuro

Definición (Problema de aprendizaje)

Sea P una distribución de probabilidad definida sobre un espacio de ejemplos $\mathcal{Z} \subset \mathbb{R}^N$, Λ un conjunto de parámetros, y $\ell : \Lambda \times \mathcal{Z} \rightarrow \mathbb{R}^+$ una función tal que para cada $\alpha \in \Lambda$ fijo, $\ell(\alpha, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}^+$ es integrable con respecto a P en \mathcal{Z} . Consideramos el funcional:

$$L_P(\alpha) \stackrel{\text{def}}{=} \int_{\mathcal{Z}} \ell(\alpha, z) dP(z) = \mathbb{E}_{z \sim P}[\ell(\alpha, z)], \quad \alpha \in \Lambda \quad (1)$$

El objetivo del aprendizaje es minimizar el funcional (1) en el caso en que la distribución de probabilidad $P(z)$ es desconocida y se dispone de una realización de una muestra aleatoria simple de P , $\mathcal{D} = \{z_1, \dots, z_m\}$.

Definición (Error empírico)

El error empírico o error de entrenamiento se define como la pérdida esperada sobre la muestra \mathcal{D} :

$$L_{\mathcal{D}}(\alpha) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \ell(\alpha, z_i) \quad (2)$$

ERM

La **regla de minimización del riesgo empírico** (ERM) busca minimizar el funcional (2) sobre el conjunto de entrenamiento dado \mathcal{D} , en lugar del error real (1).

Definición - Aprendizaje probablemente aproximadamente correcto (PAC)

Dado un problema de aprendizaje $(\mathcal{Z}, \Lambda, \ell)$, la clase de hipótesis $\mathcal{H}(\Lambda)$ es **PAC aprendible** con respecto a \mathcal{Z} y la función de pérdida ℓ , si existe una función $m_\Lambda : (0, 1)^2 \rightarrow \mathbb{N}$ y un algoritmo de aprendizaje \mathcal{A} con la siguiente propiedad:

- Para cada $\varepsilon, \delta \in (0, 1)$ y para cada distribución de probabilidad P definida sobre \mathcal{Z} , cuando se aplica el algoritmo \mathcal{A} sobre $m \geq m_\Lambda(\varepsilon, \delta)$ ejemplos i.i.d generados por P , \mathcal{A} devuelve una hipótesis $h(\cdot, \alpha)$, $\alpha \in \Lambda$, tal que

$$\mathbb{P}_{\mathcal{D} \sim P^m} [L_P(\alpha) - \inf_{\alpha' \in \Lambda} L_P(\alpha') \leq \varepsilon] \geq 1 - \delta \quad (3)$$

Definición (Dicotomías)

Sea $\mathcal{H}(\Lambda) = \{h(x, \alpha), \alpha \in \Lambda\}$ una clase de funciones hipótesis y $C = \{x_1, \dots, x_m\} \subset \mathcal{X}$. Las dicotomías generadas por $\mathcal{H}(\Lambda)$ en el subconjunto de puntos C se definen como:

$$\mathcal{H}_C(\Lambda) = \{(h(x_1, \alpha), \dots, h(x_m, \alpha)) : \alpha \in \Lambda\}$$

donde cada elemento de $\mathcal{H}_C(\Lambda)$ es un vector en $\{0, 1\}^m$.

Definición (Función de crecimiento)

La función de crecimiento de $\mathcal{H}(\Lambda)$, denotada por $\Pi_\Lambda : \mathbb{N} \rightarrow \mathbb{N}$, viene dada por

$$\Pi_\Lambda(m) = \max_{C \subset \mathcal{X} : |C|=m} |\mathcal{H}_C(\Lambda)|, \quad \forall m \in \mathbb{N}.$$

Definición (Función de crecimiento)

La función de crecimiento de $\mathcal{H}(\Lambda)$, denotada por $\Pi_{\Lambda} : \mathbb{N} \rightarrow \mathbb{N}$, viene dada por

$$\Pi_{\Lambda}(m) = \max_{C \subset \mathcal{X}: |C|=m} |\mathcal{H}_C(\Lambda)|, \quad \forall m \in \mathbb{N}.$$

Definición (Dimensión VC)

La dimensión de Vapnik-Chervonenkis (VC) de una clase de hipótesis $\mathcal{H}(\Lambda) = \{h(x, \alpha), \alpha \in \Lambda\}$, denotada por $d_{VC}(\Lambda)$, es el mayor valor de m para el cual $\Pi_{\Lambda}(m) = 2^m$. Cuando $d_{VC}(\Lambda) = \infty \forall m \in \mathbb{N}$ decimos que $\mathcal{H}(\Lambda)$ tiene dimensión VC infinita.

Desigualdad de Vapnik-Chervonenkis

Para todo $\varepsilon > 0$ se cumple que:

$$\mathbb{P}_{\mathcal{D} \sim P^m} \left[\sup_{\alpha \in \Lambda} |L_P(\alpha) - L_{\mathcal{D}}(\alpha)| > \varepsilon \right] \leq 4\pi_{\Lambda}(2m) \exp \left(-\frac{1}{8} \varepsilon^2 m \right)$$

Desigualdad de Vapnik-Chervonenkis

Para todo $\varepsilon > 0$ se cumple que:

$$\mathbb{P}_{\mathcal{D} \sim P^m}[\sup_{\alpha \in \Lambda} |L_P(\alpha) - L_{\mathcal{D}}(\alpha)| > \varepsilon] \leq 4\Pi_{\Lambda}(2m) \exp\left(-\frac{1}{8}\varepsilon^2 m\right)$$

Lema de Sauer

Sea $\mathcal{H}(\Lambda) = \{h(x, \alpha), \alpha \in \Lambda\}$ una clase de hipótesis con $d_{VC}(\Lambda) = d < \infty$. Entonces para todo $m \geq d$, se cumple que

$$\Pi_{\Lambda}(m) \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

Desigualdad de Vapnik-Chervonenkis

Para todo $\varepsilon > 0$ se cumple que:

$$\mathbb{P}_{\mathcal{D} \sim P^m} \left[\sup_{\alpha \in \Lambda} |L_P(\alpha) - L_{\mathcal{D}}(\alpha)| > \varepsilon \right] \leq 4\pi_{\Lambda}(2m) \exp \left(-\frac{1}{8} \varepsilon^2 m \right)$$

Lema de Sauer

Sea $\mathcal{H}(\Lambda) = \{h(x, \alpha), \alpha \in \Lambda\}$ una clase de hipótesis con $d_{VC}(\Lambda) = d < \infty$. Entonces para todo $m \geq d$, se cumple que

$$\pi_{\Lambda}(m) \leq \left(\frac{em}{d} \right)^d = O(m^d).$$

PAC aprendible con criterio ERM \iff dimensión VC finita

Descripción del problema

El problema del aprendizaje

Optimización

Redes neuronales artificiales

Aplicación

Trabajo Futuro

El **algoritmo de gradiente descendente** consiste en aplicar iterativamente la siguiente regla de actualización:

$$x^{(k+1)} \leftarrow x^{(k)} - \eta \nabla f(x^{(k)})$$

Teorema (Convergencia del gradiente descendente)

Sea $f : \mathbb{R}^N \rightarrow \mathbb{R}$ una función diferenciable, convexa y con gradiente Lipschitziano con constante $L \geq 0$. Sea $x^* \in \mathbb{R}^N$ el mínimo de f y $x^{(k)}$, $k = 0, 1, 2, \dots$, una sucesión de puntos en \mathbb{R}^N producida por el algoritmo de gradiente descendente con un tamaño de paso fijo $\frac{1}{L} \geq \eta > 0$. Entonces,

$$f(x^{(k)}) - f(x^*) \leq \frac{2\|x^{(0)} - x^*\|_2^2}{\eta k}.$$

El **algoritmo de gradiente descendente** consiste en aplicar iterativamente la siguiente regla de actualización:

$$x^{(k+1)} \leftarrow x^{(k)} - \eta \nabla f(x^{(k)})$$

Teorema (Convergencia del gradiente descendente)

Sea $f : \mathbb{R}^N \rightarrow \mathbb{R}$ una función diferenciable, convexa y con gradiente Lipschitziano con constante $L \geq 0$. Sea $x^* \in \mathbb{R}^N$ el mínimo de f y $x^{(k)}$, $k = 0, 1, 2, \dots$, una sucesión de puntos en \mathbb{R}^N producida por el algoritmo de gradiente descendente con un tamaño de paso fijo $\frac{1}{L} \geq \eta > 0$. Entonces,

$$f(x^{(k)}) - f(x^*) \leq \frac{2\|x^{(0)} - x^*\|_2^2}{\eta k}.$$

Otras variantes del gradiente descendente usadas en el aprendizaje: gradiente descendente estocástico, gradiente descendente con momento, RMSProp, Adagrad, *Adam*, etc.

Descripción del problema

El problema del aprendizaje

Optimización

Redes neuronales artificiales

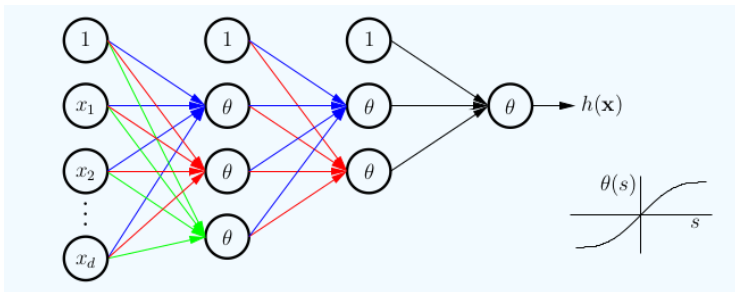
Aplicación

Trabajo Futuro

Definición (neurona)

Sea $w \in \mathbb{R}^d$, $d \in \mathbb{N}$, y $\theta : \mathbb{R} \rightarrow \mathmathbb{R}$ una función escalar, conocida como *función de activación*. Una neurona se define como una función $f : \mathbb{R}^{2d} \rightarrow \mathbb{R}$ dada por

$$f(x, w) = \theta \left(\sum_{i=1}^d w_i x_i \right), \quad \forall x, w \in \mathbb{R}^d.$$



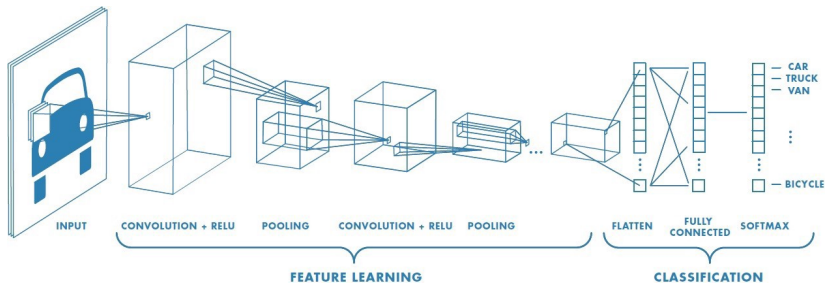
- ▶ Gran poder de aproximación: **Teorema de Aproximación Universal**.
- ▶ Tienen dimensión VC finita en general. Por ejemplo:

Teorema (Bartlett, Harvey, Liaw y Mehrabian, 2017)

Sea una red neuronal prealimentada con T capas, pesos w y una única neurona de salida con función de activación $\mathbb{1}_{[0,\infty)}$. Si todas las neuronas en las capas ocultas usan funciones de activación lineales a trozos con un número finito de trozos, llamando $\mathcal{H}_{T,d,\theta}$ a la clase de funciones reales implementadas por dicha red, entonces existen constantes c, C tales que

$$c \cdot wT \log(w/T) \leq d_{VC}(\mathcal{H}_{T,d,\theta}) \leq C \cdot wT \log(w).$$

Tipo especial de redes neuronales prealimentadas en las que algunas capas (*capas convolucionales*) implementan una operación de **convolución**.



0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114				

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328			

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26		

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158
53				

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158
53	266			

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158
53	266	-61		

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158
53	266	-61	-30	

0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158
53	266	-61	-30	344
403	116	-47	295	

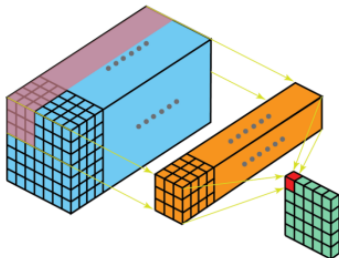
0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

Kernel

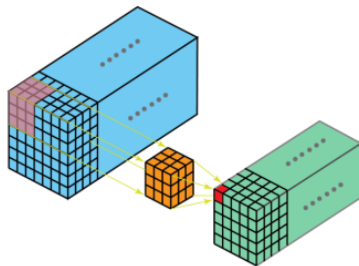
0	-1	0
-1	5	-1
0	-1	0

114	328	-26	470	158
53	266	-61	-30	344
403	116	-47	295	244
108	-135	256	-128	344
314	346	279	153	421

- ▶ Los filtros son tridimensionales y se aplican en tres dimensiones.
- ▶ Diferente a la convolución 2D multicanal.

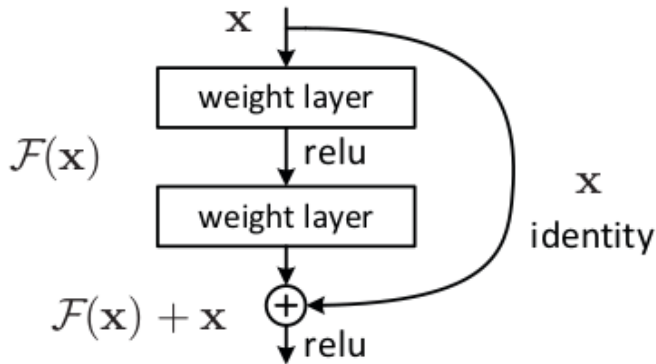


(a) Multi-channel 2D convolution



(b) 3D convolution

- ▶ Surgen como solución al problema de *degradación*.
- ▶ Se incluye una conexión directa (*shortcut connection*) entre la entrada y la salida de las capas que aplica la función identidad.



Descripción del problema

El problema del aprendizaje

Optimización

Redes neuronales artificiales

Aplicación

Trabajo Futuro

Usamos distintas bases de datos públicas.

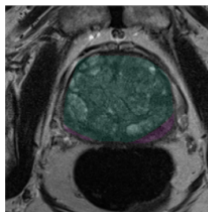
- ▶ Segmentación de las lesiones de Cap: 375 pacientes, 185 con al menos un tumor maligno y 190 sanos → balanceo.
 - ▶ **ProstateX** (incluye lesiones benignas).
 - ▶ **I2CVB**.
 - ▶ 3T Siemens escáner.
 - ▶ 1.5T GE (General Electric) escáner.
 - ▶ **Prostate158**.

Las imágenes difieren en dimensionalidad y en tamaño de voxels.

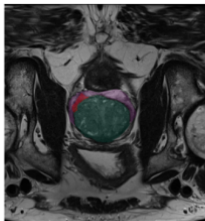
Usamos distintas bases de datos públicas.

- ▶ Segmentación de la anatomía de la próstata:
37 imágenes con solo las segmentaciones de la *próstata*
92 imágenes con las segmentaciones de las *zonas*.
 - ▶ **PROMISE12** (sólo incluye máscaras de la próstata).
 - ▶ **Medical Segmentation Decathlon (Tarea 5).**
 - ▶ **NCI-ISBI.**

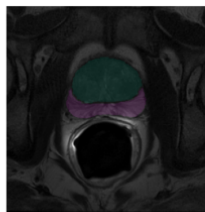
Las imágenes difieren en dimensionalidad y en tamaño de voxels.



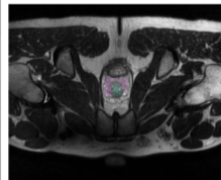
(a) Prostate158



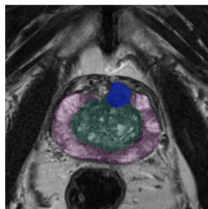
(b) ProstateX



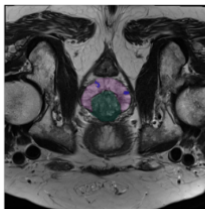
(c) I2CVB GE scanner



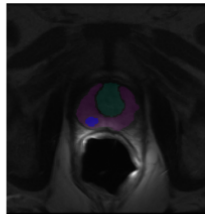
(d) I2CVB Siemens scanner



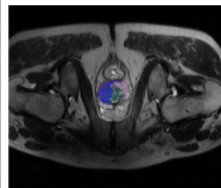
(e) Prostate158



(f) ProstateX



(g) I2CVB GE scanner



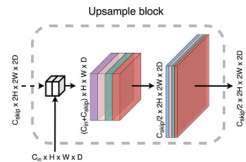
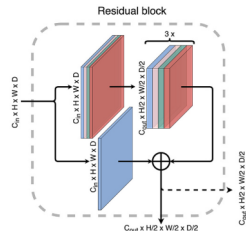
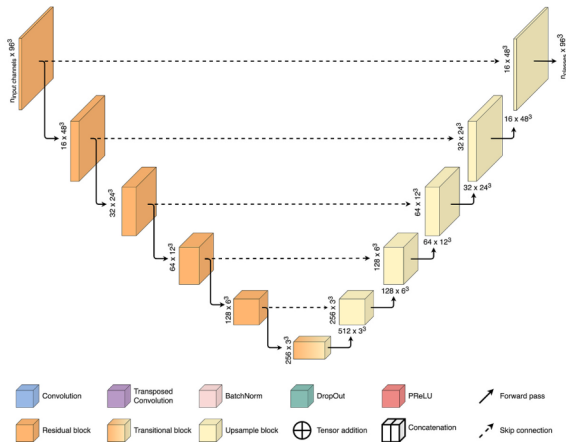
(h) I2CVB Siemens scanner



Solo usamos la modalidad de **imagen por resonancia magnética potenciada en T2**. ¿Por qué?

Solo usamos la modalidad de **imagen por resonancia magnética potenciada en T2**. ¿Por qué?

- ▶ El tiempo necesario para adquirir varias modalidades es muy alto.
- ▶ Es necesaria una operacion de registro (alineación) de las distintas modalidades.
- ▶ Es la modalidad con mayor resolución.

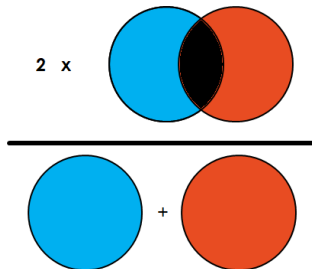


Medida de superposición entre dos muestras.

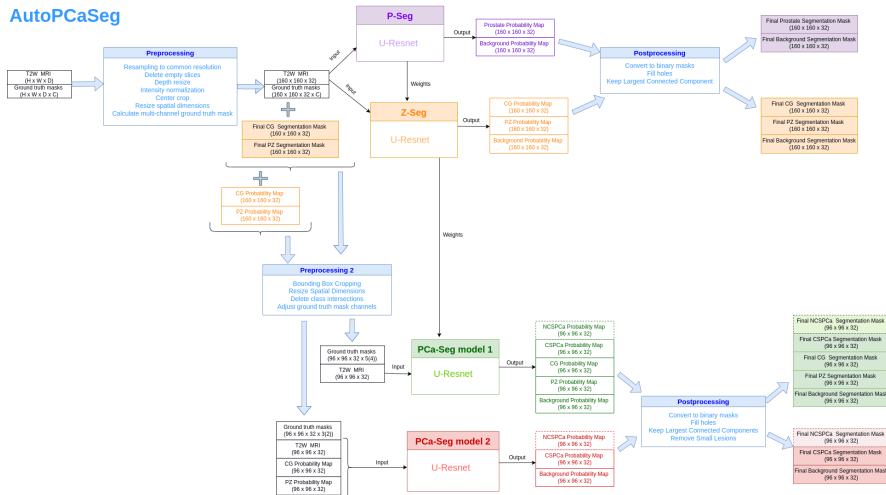
Dados dos conjuntos A y B , el coeficiente Dice se define como:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}. \quad (4)$$

Función de pérdida Dice: $1 - DSC$



AutoPCaSeg



- ▶ Validación cruzada con 3 plieges
 - ▶ **P-Seg**: valor medio de DSC de 0,9049
 - ▶ **Z-Seg**: valor medio de DSC de 0,8506 y 0,7283 para la GC y ZP, respectivamente.
- ▶ Test
 - ▶ **P-Seg**: valores de DSC oscilan entre 0,858 y 0,917
 - ▶ **Z-Seg**: valores de DSC en el rango de 0,67 a 0,877 para GC y 0,637 a 0,7731 para ZP.

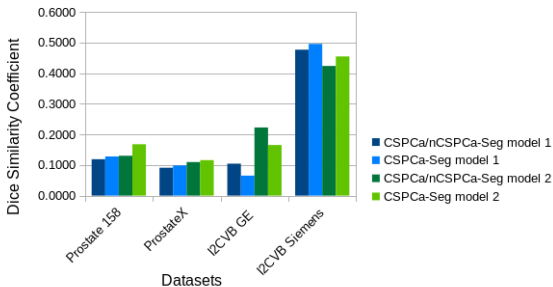
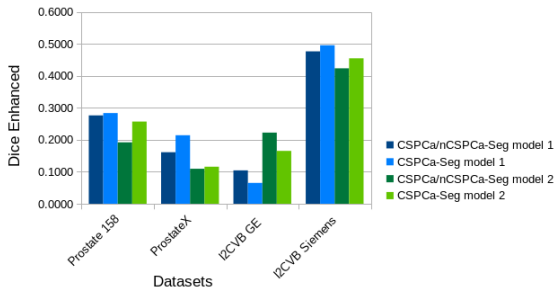
Resultados similares al estado del arte!

► Validación cruzada con 3 plieges:

Mejor DSC para lesiones de CaP **clínicamente significativas**:
0,16.

► Test:

Segmentación de **lesiones benignas** en ProstateX:
el modelo 2 PCa-Seg alcanzó un DSC de 0,118.





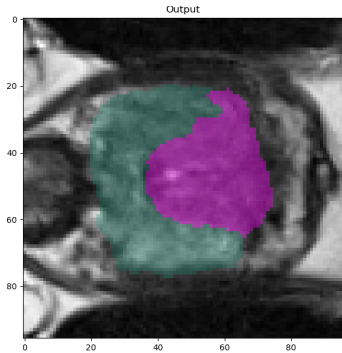
Conclusiones:

- ▶ Diferencias entre modelos no significativas.
- ▶ Diferencias entre datasets sí significativas → inevitables.
- ▶ Resultados por debajo del estado del arte. ¿Por qué?

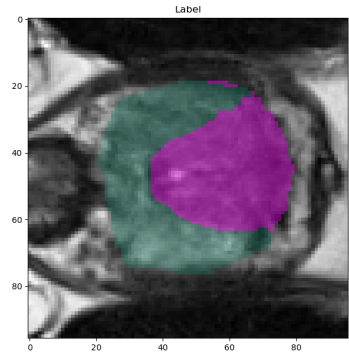
Conclusiones:

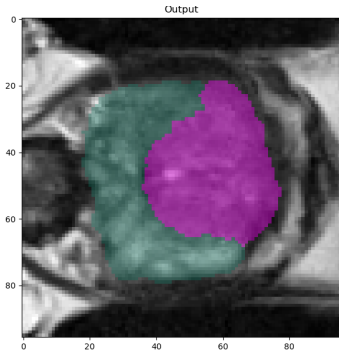
Resultados por debajo del estado del arte. ¿Por qué?

- ▶ Nuestro modelo es simple.
- ▶ Usamos una única modalidad de IRM.
- ▶ Usamos datasets heterogéneos para entrenar.
- ▶ Díficil comparar con estudios previos.

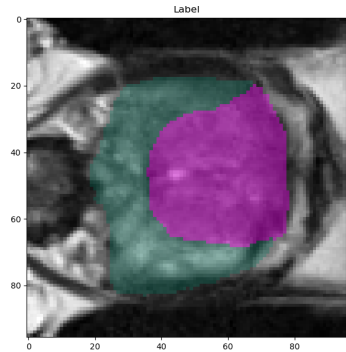


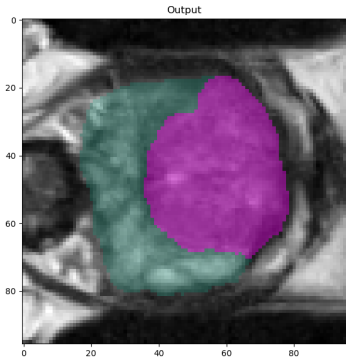
Patient 31, slice 9



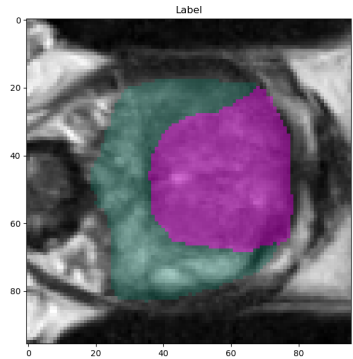


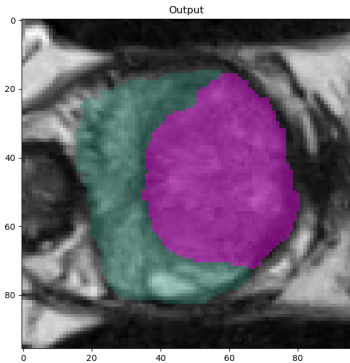
Patient 31, slice 10



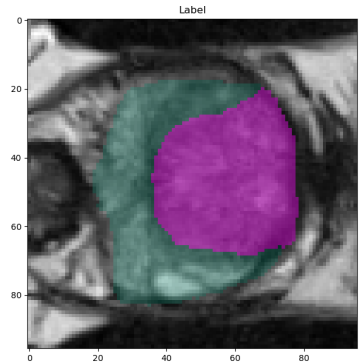


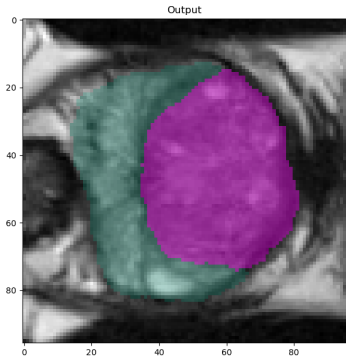
Patient 31, slice 11



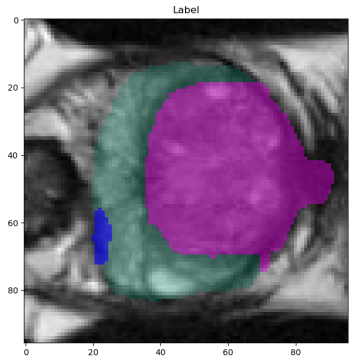


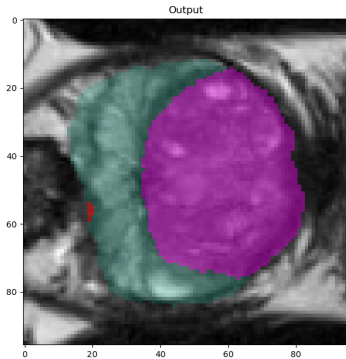
Patient 31, slice 12



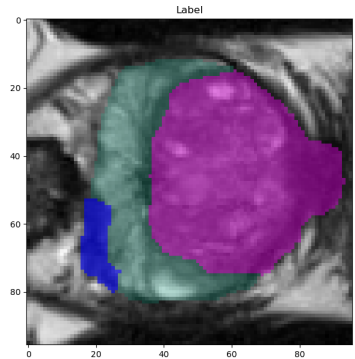


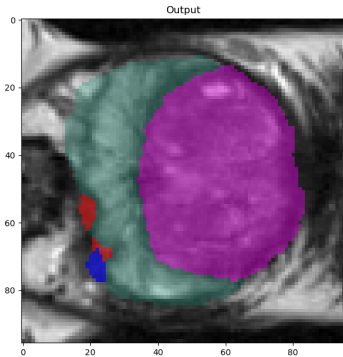
Patient 31, slice 13



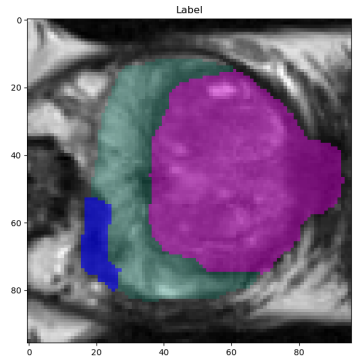


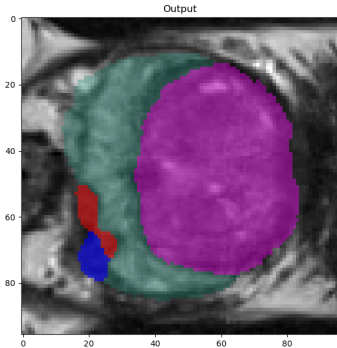
Patient 31, slice 14



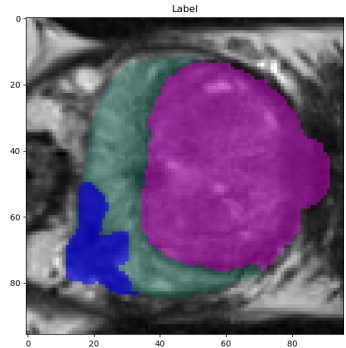


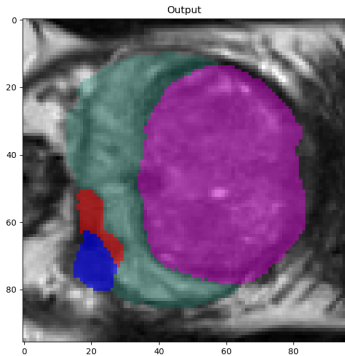
Patient 31, slice 15



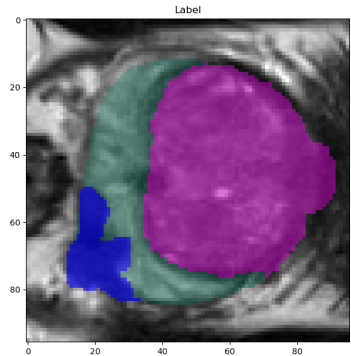


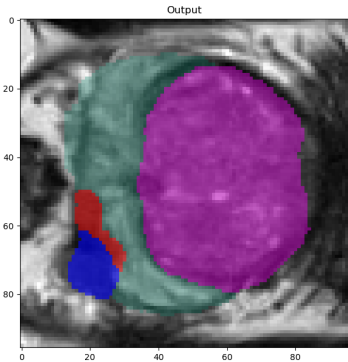
Patient 31, slice 16



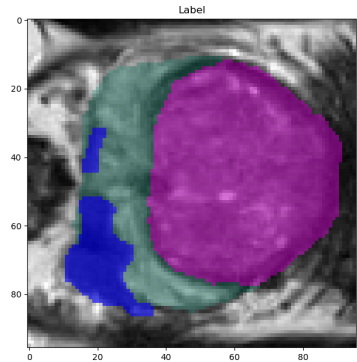


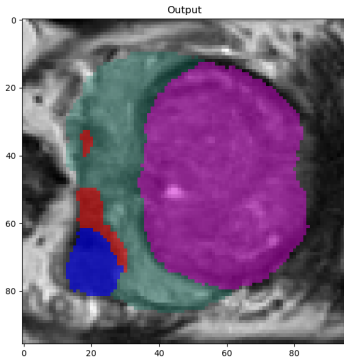
Patient 31, slice 17



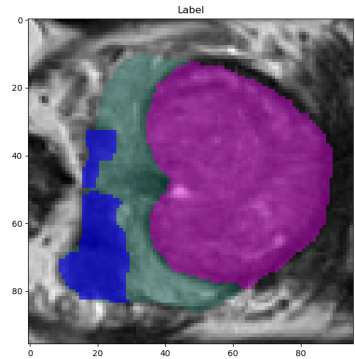


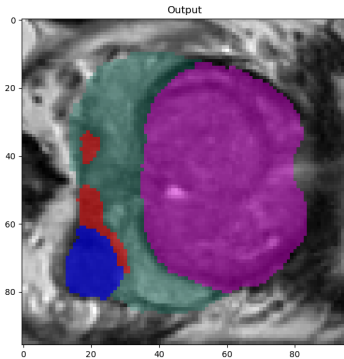
Patient 31, slice 18



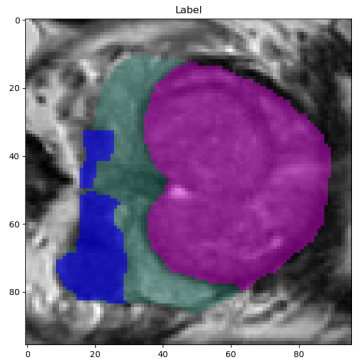


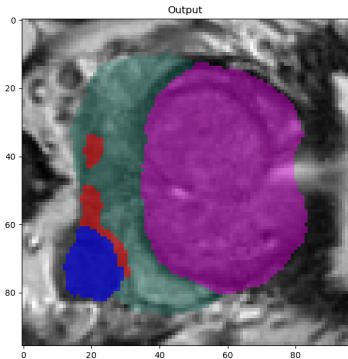
Patient 31, slice 19



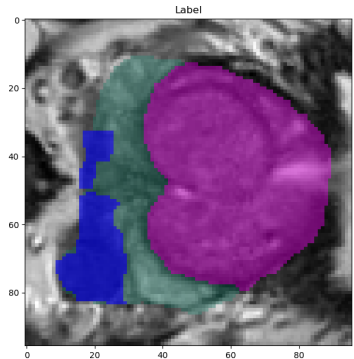


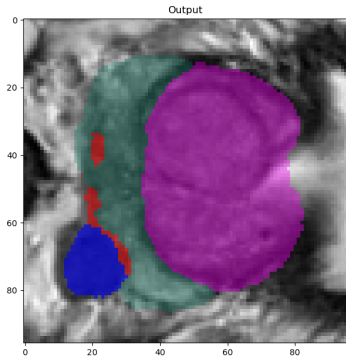
Patient 31, slice 20



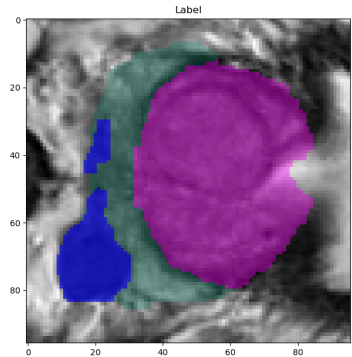


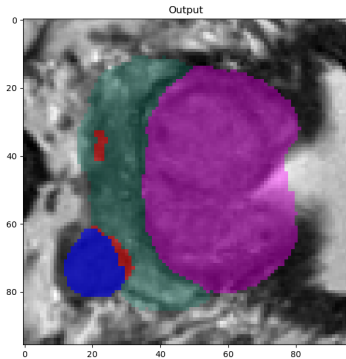
Patient 31, slice 21



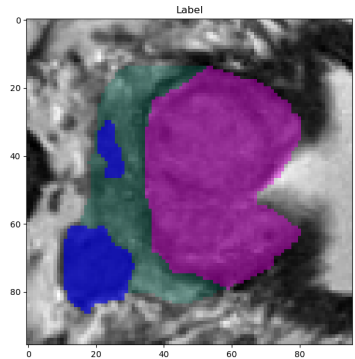


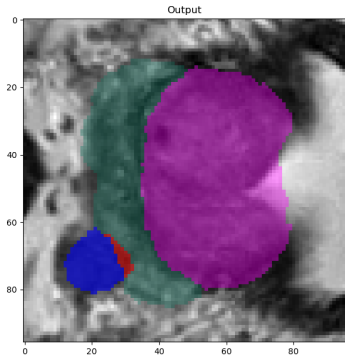
Patient 31, slice 22



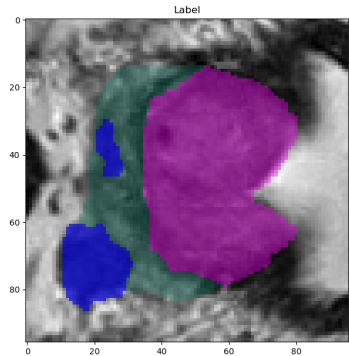


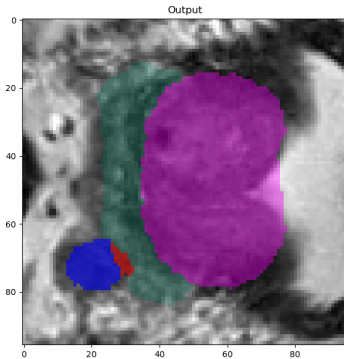
Patient 31, slice 23



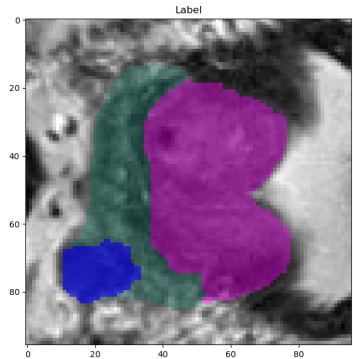


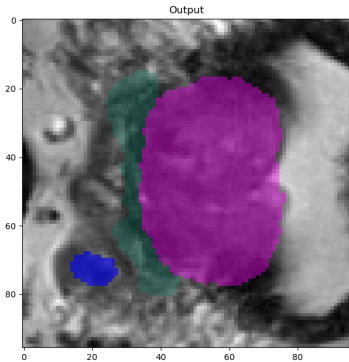
Patient 31, slice 24



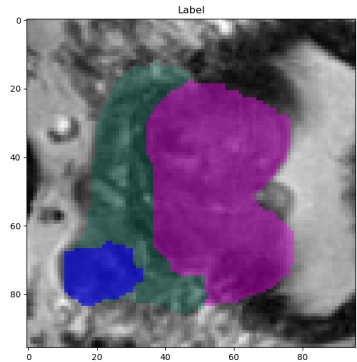


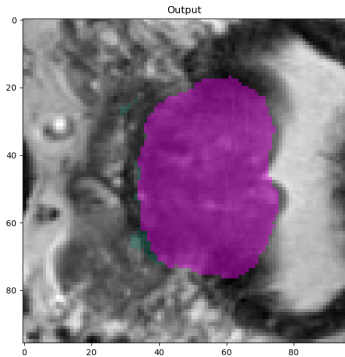
Patient 31, slice 25



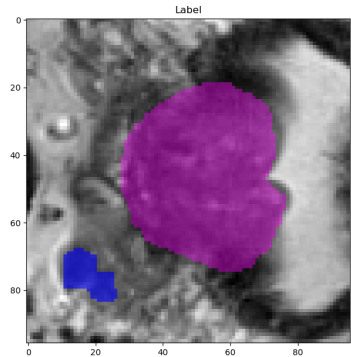


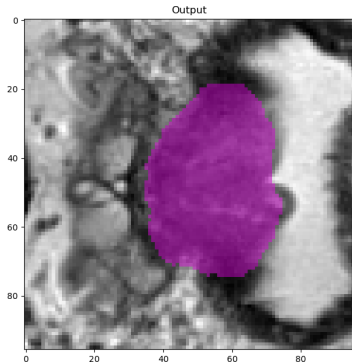
Patient 31, slice 26



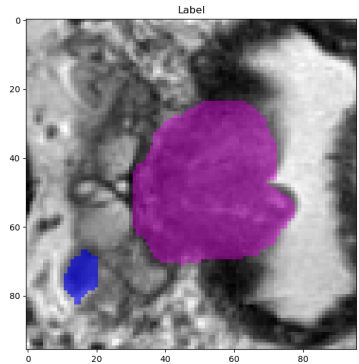


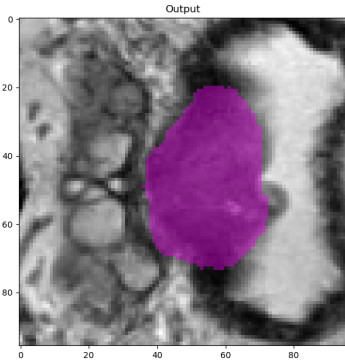
Patient 31, slice 27



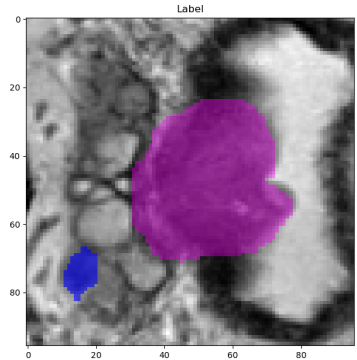


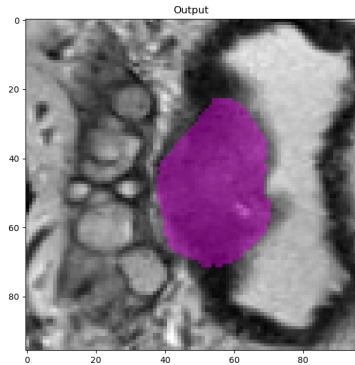
Patient 31, slice 28



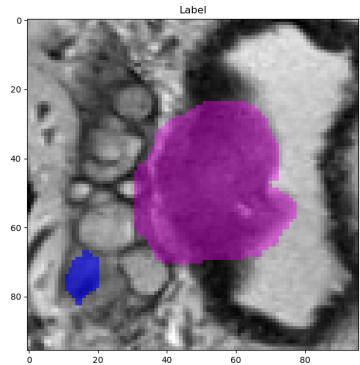


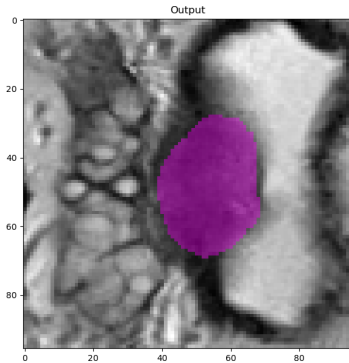
Patient 31, slice 29



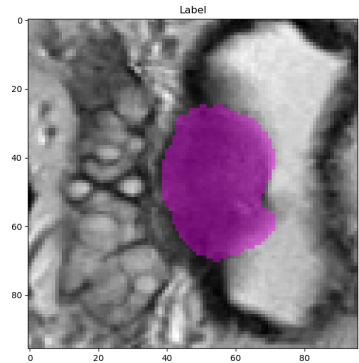


Patient 31, slice 30





Patient 31, slice 31



Descripción del problema

El problema del aprendizaje

Optimización

Redes neuronales artificiales

Aplicación

Trabajo Futuro



Queda mucho por mejorar...

Queda mucho por mejorar...

- ▶ Inclusión de diferentes modalidades IRM.
- ▶ Arquitecturas de redes neuronales más complejas: mecanismos de atención, redes generativas adversarias...
- ▶ Usar tamaño de kernel y de stride distinto en cada dimensión.
- ▶ Ensemble de varios modelos.
- ▶ Arquitecturas 2D.
- ▶ Entrenar con un solo dataset homogéneo.
- ▶ Conseguir datasets más grandes.

Etc.

¡GRACIAS POR SU ATENCIÓN!

